Self-Other Agreement on Performance Ratings as a Predictor of Individual Longitudinal Outcomes and Future Agreement

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Melissa Sue Sharpe

IN PARTIAL FULTILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Paul R. Sackett, Adviser

December 2017

Acknowledgements

I would first like to acknowledge those who helped me in my academic career: Paul Sackett, for being an ever-patient, knowledgeable, and supportive adviser—it would not have been possible without your help and guidance; my committee members (Nathan Kuncel, Aaron Schmidt, Lou Quast), for providing me candid feedback on both my dissertation and other projects throughout my career; my collaborators (Filip Lievens, R. Lee Penn, Bill Arnold), for providing me with interesting projects to stretch my skills and build my research interest; and my fellow students (Jeff Jones, Rachael Klein, Sarah Semmel, Allen Goebl, Chris Huber, Laura Johnson, Kyle McNeal, Amanda Kopydlowski, Brenton Wiernik, Brittany Marcus-Blank, Amy Shu, Tetsuhiro Yamada, Martin Yu, Oren Shewach, Win Matsuda), for serving as a sounding board for research ideas and being there to support me as a friend and peer—you made the experience of graduate school great. I would also like to acknowledge those that supported me socially: my family (Mom, Dad, Amanda, Matthew, Jessica), for always being there to listen to my problems and make me feel loved; and my friends (Phil Norgard, Mackenzie Voeller, Davey Maxam, Lauren Maxam, Mike Schumacher, Mike Francomb, Corey Krizak), for acting as a reprieve from my studies whenever I needed it. Also, to all persons unnamed that feel like they should be included but due to my carelessness I have neglected—I'm sorry, and thank you. Finally, I must acknowledge my husband, Earl Sharpe, for your patience while I worked on my studies, your support through the good times and the bad, and your ever-present and loving companionship.

Dedication

To all that supported me throughout my life and helped me get to where I am today.

Abstract

Performance appraisals have come under fire recently in the field of Industrial-Organizational Psychology, sparking some debate as to whether the field should continue the practice. The primary goal of this dissertation is to suggest that performance appraisals, within the lens of feedback, are a valuable tool and have some meaningful implications for individuals within organizations. The results from a 5-year archival longitudinal study suggest that (1) individuals tend to disagree initially from their manager's rating of their performance, but converge with time; (2) the initial and longitudinal agreement in these ratings predicts individual outcomes (e.g. salary, organizational level, promotions) in predictable ways; and (3) that participation in another form of feedback procedure (i.e. a 360º feedback program) does not impact individual performance rating trajectories, but does influence the manager's ratings of that performance in negative ways.

Table of Contents

List of Tables

List of Figures

Feedback as an Organizational Intervention

Employee performance evaluations are a ubiquitous and recurring management practice. They can serve many purposes: they are used strategically to link employee behavior and organizational goals; they also communicate important job-relevant information to employees; they often serve as the basis for employment decisions (e.g., promotion, training, or discipline); they are used as criteria in research; they can also influence employee's personal and professional development; they can also be used to evaluate the organization's effectiveness (Casio & Aguinis, 2005). It seems logical, then, that much research and theoretical attention has been paid to the performance appraisal process in both its quality of measurement and effectiveness in changing behavior.

However, some debate has recently arisen in the Industrial-Organizational (I-O) Psychology field over whether performance ratings should be abandoned altogether (see Adler, et al., 2016 for an overview). Researchers in favor of discarding performance ratings suggest that they are "a failed experiment", citing its inability to improve accuracy with different rating scales (Landy & Farr, 1980) or rater training (Murphy & Cleveland, 1995), that irrelevant contextual effects play too large a role (e.g., political aspects), or that they often suffer from adverse impact amongst racial minorities (McKay & McDaniel, 2006). Others (e.g., Woehr & Roch, 2016) suggest that performance ratings do have value within the performance management process—"so while performance management in organizations may be a messy, poorly managed, and poorly implemented process, we should be cautious not to lay the blame on the quality of performance ratings" (Woehr & Roch, 2016, p. 360). Indeed, as Adler et al. (2016) put it: "the science of behavioral change… tells us that some discomfort, some feedback on the gap between

the ‘now’ state and the desired future is required to stimulate behavior change and development” (p. 235).

Murphy and Cleveland (1995) discussed the role of performance appraisal as a form of feedback in-depth—they suggested that it may be a good idea for organizations to implement two separate systems to fulfill the multiple goals of performance reviews: “it may be a grave mistake to let the administrative side of appraisal interfere with the use of appraisal as a feedback tool” (p. 91). However, whether or not the performance appraisal is specified or designed to provide feedback to the employee does not impact its function as a feedback mechanism—it intrinsically provides some cues to the employee about their relative standing in the organization and their level performance as perceived by some other individual (typically the supervisor).

The goals of this dissertation were: (1) to provide a background on and history of the theories of feedback, (2) to cite evidence for feedback’s effectiveness at changing behavior in organizational settings, (3) to discuss the relevant measurement considerations in the evaluation of job performance, and (4) to assess how self-other agreement plays a role in the appraisal process—complete with a discussion of its own models and findings within that niche of the performance appraisal literature. To clarify, the introduction focuses on the feedback function of ratings: particularly those made in the multisource feedback and performance appraisal literatures. Studies of ratings made in other domains (e.g., interviews, job applications, assessment centers) were avoided unless research in performance appraisal was sparse. This review will not touch on feedback-seeking behavior (see Anseel, Beatty, Shen, Lievens, & Sackett, 2015 for a

review), or on impression management tactics in performance appraisal[1] (see Villanova & Bernardin, 1989 for a review) due to the expansive nature of both sets of literature and their tangential relevance to the goals outlined above. Following the literature review is a discussion of the present study and its contributions to the literature.

Theories of Feedback

The theoretical history of feedback in organizational settings has its bases in early experimental psychology and is particularly steeped in the behaviorist tradition—researchers interested in how feedback could influence future behavior and learning developed theories of human behavior on the basis of animals' abilities to learn how to complete experimental tasks in the laboratory. Later, theorizers applied these experimental findings to organizational settings—particularly to understand feedback's effect on job performance—and adjusted their frameworks to suit their empirical findings and to direct future research. A history of these theories is outlined below.

The Law of Effect. Thorndike's (1913) law of effect is arguably the most fundamental theory in understanding feedback. This law asserts that behaviors that produce a "satisfying effect" (i.e., reinforcement) in a certain situation are more likely to reoccur in that situation, while the reverse is true of behaviors that produce a "discomforting" effect. (i.e., punishment). Thus, performance should be improved regardless of the sign of feedback because positive feedback should increase correct behaviors and negative feedback should decrease incorrect ones. The law tended to hold very well in behaviorists' experiments with animals. However, the principle did not

[1] However, the role of organizational politics will be discussed; these research areas have a large amount of overlap and therefore there may be expectations to discuss both.

always hold in human learning environments. Annett (1969) outlined several empirical findings which subverted this rule, and declared that the law of effect "has been rejected on both empirical and logical grounds" (p.169). Kluger and DeNisi (1996) similarly noted that the law of effect "has the advantage of parsimony, but it is too broad to explain the empirical complexities associated with FI [Feedback Interventions]" (p. 259).

Knowledge of Results. Harry Kay, in his foreword to Annett's (1969) book on feedback, noted the effect of technology on our understanding of behavior: "The computer age is underway. This orientation which has been so much influenced by cybernetics and information theory has permeated our whole approach to human learning." Indeed, early understanding of the influence of feedback on human behavior was very reminiscent of machine and computerized learning. Annett (1969) outlined the progression of thought in this area: from simple concepts like "stimulus leads to a response" (i.e., S → R), to servo-mechanisms and reinforcement theory. The focus of his book, and what later became one of the bases of the feedback literature, was on Knowledge of Results (KR). He noted the difference between intrinsic KR—"that which is normally present and is not often subject to experimenter manipulation"—and extrinsic KR—"feedback being supplied by the experimenter or specially adapted by him" (Annett, 1969 p.26). As such, feedback was believed to have served three functions: an informative function, a reinforcing function, and an incentive function.

However, Annett (1969) notes that KR researchers tended to "assert that KR has all three properties but decline to say how they are mixed or how they can be disentangled" (p. 37). He later concluded that the effects of each function could not, and should not, be isolated—"our main task has been to show that all the supposedly different

functions of KR can be derived from the properties of feedback systems without recourse to a 'mixed' theory" (p.160). Annett (1969) further reduced the issue: "The informative value of KR is seen in terms not only of the information content of the 'results' but also in relation to the kind of transformation rule the learner is using" (p. 169). Thus, the focus of KR researchers moved away from delineating the relative effectiveness of feedback along these three variables, but still lacked an organizing framework.

Feedback in Organizations. Ilgen, Fisher, and Taylor (1979) took note of this issue, particularly within the organizational feedback literature. At the time, the literature on feedback in organizational settings was somewhat disjointed—generalizations about the effects of feedback were difficult to make. They suggested that this was due mostly to two factors: (1) that feedback is multifaceted and complex; one operationalization of feedback did not exist, and (2) that there was no guiding theoretical foundation to direct research in an efficient manner. They created a model in order to address the second issue (see Figure 1). However, they do note that their model was developed as an organizing framework rather than a theory, per se: "it [their model] is meant to serve primarily as a vehicle for organizing the discussion of feedback rather than to be a well-developed theory of feedback" (p. 352). Naylor, Pritchard, and Ilgen (1980) later expanded the model to encompass all organizational behavior, but did not get much attention from researchers. The original (1979) model often is cited as a theory in its own right—Fedor (1991) notes that this model is the most frequently used.

Control Theory. Another highly influential theory in the feedback literature is control theory (Carver & Scheier, 1981), the basis of which asserts that behavior can be understood as a series of negative feedback loops: standards of behavior and actual

behavior are compared—any discrepancy between the two encourages the individual to reduce the discrepancy through his or her future behavior. A discrepancy between feedback and standards can be reduced in four ways: by changing behavior to reduce future discrepancy, by changing the standard to match current feedback, by rejecting the feedback outright, or by exiting the situation, either physically or mentally. Taylor, Fisher, and Ilgen (1984) merged their earlier model (Ilgen, Fisher, and Taylor, 1979) with control theory as a method of understanding individual's reactions to performance feedback. Control theory, to the authors, explained the structure of standard hierarchies and emphasized important aspects of the feedback process (e.g., congruence between feedback dimensions and behavioral standards). However, other researchers did not find the value of control theory in understanding human behavior.

Goal Setting Theory. Goal setting theory (Locke & Latham, 1990; 2002) is, at its surface (in terms of how feedback is used by the recipient), very similar to control theory. Both theories suggest that a difference between feedback and standards influence behavior. Similarly, the four behavioral options for responses to discrepancies in goal setting theory mimic those in control theory: an individual can strive to attain the goal, change the goal, reject the feedback, or abandon commitment to the goal. However, control theorists suggest that the individual is motivated by reducing the discrepancy, whereas goal setting theorists suggest that the individual is motivated by achieving the goal. This difference is somewhat subtle, but stems from the difference in foundation of the two theories: control theory developed as a result of mechanical systems building, of machine building. Goal theory, on the other hand, arose after an empirical review of human behavior. Locke (1991) provided a detailed comparison of the two theories, and

outlined many reasons why he felt that control theory was not applicable to understanding human motivation—one of the major arguments suggests that humans do not simply attempt to reduce discrepancies between goal states and current states, but in fact are at times discrepancy-producing. That is, "discrepancy reduction actually is a *consequence* of goal-directed behavior, not its cause" (Locke, 1991 p. 13, emphasis original). However, while goal setting theory is *related* to feedback—and explicitly incorporates it as an element—it is not explicitly a theory about feedback as an organizational intervention.

Feedback Intervention Theory. Feedback Intervention Theory (FIT; Kluger & DeNisi, 1996) was developed to incorporate elements of pre-existing theories related to feedback in order to explain the inconsistent findings of effectiveness for organizational feedback interventions (FIs) on increasing job performance. The five basic tenets of this theory are: (1) that behavior is regulated by comparison of feedback to goals or standards, (2) these goals and standards are arranged hierarchically, (3) attention is limited, therefore only discrepancies between standards and performance that are given attention will dictate future behavior, (4) attention is typically directed at a moderate level of the goal hierarchy, and (5) the main purpose of FIs is to change the locus of attention and therefore influence behavior (Kluger & DeNisi, 1996, p. 259). The first four arguments come almost directly from control theory and goal setting theory (and each have their own empirical support), but the last one, they argue, is unique to FIT and critical for understanding the relationship between feedback and performance.

Their theory reduces down to the supposition that feedback can orient the attention of the recipient to either themselves ("meta-task goals"), the task at hand, or to

certain components of the task at hand, and these "goals" are arranged hierarchically. Feedback that directs attention toward meta-task goals was posited to attenuate the relationship between feedback and performance, particularly through a diversion of cognitive resources away from performance and toward affective responses. Feedback that directs the recipient to the task at hand would be associated with increased motivation, and therefore increased performance. Feedback that directs attention to task details would increase learning, and therefore performance. The relationship between feedback characteristics ("cues") and the level of the goal hierarchy to which it directs the recipient would be influenced by both situational and personality variables. Similarly, the relationship between attention to the goal and subsequent performance would be influenced by task characteristic variables. A graphic representation of this model can be found in Figure 2.

Evidence for Effectiveness

Several reviews of the feedback literature have been conducted over its history to assess whether or not it is effective at changing behavior (e.g., reducing absenteeism, increasing job performance). Early reviewers simply tallied the number of studies on the topic according to the direction of their result. Balcazar, Hopkins, and Suarez (1985) reviewed the feedback literature published in four major journals from 1975 to 1985: *Academy of Management Journal*, *Journal of Applied Behavior Analysis*, *Journal of Applied Psychology*, and *Journal of Organizational Behavior Management*. They found that in most studies, feedback did not uniformly improve performance—only 28% of the 129 instances of feedback interventions that they found consistently increased performance. Alvero, Bucklin, and Austin (2001) reviewed the same journals from 1985

to 1998 and found similar results: 47% of their located studies of feedback consistently increased performance.

Other reviewers have focused on the effectiveness of specific intervention systems which incorporate a feedback element. Management by objectives (MBO) is a system of performance management which has three main components: goal setting, participative decision making, and objective feedback (Drucker, 1976). Rodgers and Hunter (1991) found positive results of an MBO system on productivity in 97% of their located studies—a substantial positive effect (d = 0.42) was also found in the 12 studies that used performance rating data. Pritchard, Harrell, Diaz Granados, and Guzman (2008) conducted a meta-analysis on studies of the Productivity Measurement and Enhancement System (ProMES)—a method which defines organizational objectives and indicators of progress toward the objectives, followed by feedback reports and a feedback meeting. They found that this intervention was overall effective at increasing performance (d = 1.16), particularly after the feedback portion of the intervention had begun.

Smither, London, and Reilly (2005) conducted a meta-analysis on longitudinal studies of performance feedback—specifically longitudinal studies of multisource feedback. Although their located study size was small (k = 21), they found modest effect sizes for both upward feedback (i.e., from a direct report to a manger; d = 0.24) and multisource feedback (d = 0.24). The most comprehensive, and most definitive, meta-analysis of the feedback literature was conducted by Kluger and DeNisi (1996). They included only studies which examined the effects of performance feedback alone, included a control group (or a quasi-control group), measured performance, and had 10 or

more participants—resulting in 131 studies yielding 607 effect sizes calculated from 12,652 participants. They found a reasonably high uncorrected effect size (d = 0.41) for feedback interventions, but the distribution of effect sizes was highly variable (σ = 0.97); 33% of the effect sizes were negative, even after exclusion of a set of potentially questionable studies by one researcher. Therefore, the search for moderators was necessary.

Moderators. Researchers have also examined the effects of various moderating variables on the relationship between feedback interventions and their ability to change behavior. Some of these moderators are characteristics of the feedback itself (e.g., source, valence, and frequency) while others are characteristics of the environment.

Combination. In many organizational settings, performance feedback is used in conjunction with other interventions such as goal setting and behavioral consequences. The studies reviewed by Balcazar, Hopkins, and Suarez (1985) that used other interventions in addition to feedback had more consistently positive results than those that used feedback alone. Alvero, Bucklin, and Austin (2001) replicated these results in their review of the literature in the 1990s. Kluger and DeNisi (1996) found a moderator effect of goal setting use in addition to feedback such that goal setting improves performance (d = 0.51) to a greater degree than feedback without it (d = 0.30).

These results have been confirmed in more recent investigations as well. For example, in a confirmatory factor analysis, Jawahar (2010) found evidence for a direct effect of feedback used in conjunction with goal setting on job performance of employees in a software company. Goal setting is not the only intervention that has been studied: Anseel, Lievens, and Schollaert (2009) examined the effects of reflection on subsequent

task performance. They asked some participants to reflect on their previous performance by writing down what they thought they had done well and what they had done poorly (along with behavioral examples of each). They found that feedback increased performance, reflection without feedback did not improve performance, but reflection coupled with feedback increased performance over and above feedback alone.

Source. The source of the feedback also has been found to impact the effectiveness of the feedback. Balcazar (1985) found that studies with feedback from supervisors were more likely to have consistent positive effects (50%) than studies with feedback from the researchers (33%) or from participants themselves (21%). Similar results (59% in supervisory feedback, 50% in researcher generated feedback) were found by Alvero, Bucklin, and Austin (2001). In contrast, Smither, London, and Reilly (2005) found that the longitudinal effects of feedback were strongest for feedback from direct reports and peers (both $d = 0.15$, corrected for unreliability and sampling error), weak for feedback from supervisors ($d = 0.07$), and basically non-existent for feedback from the self ($d = 0.03$).

Other researchers have investigated characteristics of the source other than their relative organizational standing (i.e., peer versus direct report versus supervisor). Fedor, Davis, Maslyn, and Mathieson (2001) examined the effects of the French and Raven (1959) bases of power of the rater on performance improvement following feedback. They found a positive association between performance improvement and managerial possession of both expert power and referent power and a negative relationship between possession of reward power and performance improvement. Similarly, Kinicki, Prussia,

Wu, and McKee-Ryan (2004) found that the perceived credibility of the feedback source was positively related (through other mediating variables) to job performance.

Valence. Whether the feedback is generally positive or generally negative has been suggested as a potential moderator of its effectiveness. In their entire sample of studies, Kluger and DeNisi (1996) found that the sign of the feedback was correlated with effect size (r = 0.24), but after removing suspected studies[2] the effect went away (r = -0.01). Studies conducted after data collection by Kluger and DeNisi (1996) found that self-ratings of performance increased after receipt of positive feedback (Atwater, Roush, & Fischthal, 1995; Bailey & Austin, 2006). Evidence also suggests that supervisory ratings of later performance is positively associated with receipt of positive feedback (Kinicki, Prussia, Wu, & McKee-Ryan, 2004; Zheng, Diaz, Jing, & Chiaburu, 2015).

Researchers have found that negative feedback has produced decreases in self-ratings but increases in follower ratings (Atwarer, Roush, & Fischthal, 1995). Also, Brutus, London, and Martineau (1999) found that receiving negative feedback was associated with increased goal-setting behavior. Others found that valence of the feedback had no effect on future performance (Atwater, Waldman, Atwater, & Cartier, 2000; Vancouver & Tischner, 2004) or on engagement in developmental activity (Bailey & Austin, 2006). However, research has also shown that positive feedback is seen as more accurate (Brett & Atwater, 2001) and is generally more accepted by the ratee (Facteau, Facteau, Schoel, Russel, & Poteet, 1998) than is negative feedback.

[2] Kluger and DeNisi (1996) decided to include variations of their analyses with excluded studies conducted by a researcher named Mikulincer—their studies provided estimates that were outliers and were therefore considered suspect by the authors.

Frequency. Salmoni, Schmidt, and Walter (1984) were among the first to conclude that more feedback was more effective than less feedback—particularly for tasks without intrinsic feedback (i.e., tasks for which progress cannot be marked without external consultation). However, much of this research was conducted on the performance of physical or motor tasks, usually in the laboratory. Reviews of studies of the frequency of job performance feedback have been less conclusive. Balcazar, Hopkins, and Suarez (1985) found an even split in the number of studies that found consistent effects and mixed effects of feedback in both studies of daily and weekly feedback. Alvero, Bucklin, and Austin (2001) had even less conclusive results, finding considerable evidence for effectiveness of feedback in studies of daily and monthly feedback, but not in those of weekly feedback.

Meta-analyses in this area have also yielded inconclusive results. After removal of suspect studies, Kluger and DeNisi (1996) found a modest, but significant, positive correlation between the frequency of feedback and its effectiveness ($r = 0.15$). However, they conclude that the effect is likely an artifact; when feedback frequency was dichotomized (by taking studies from the top and bottom quartiles of feedback frequency) less frequent feedback was found to be more effective ($d = 0.39$) than more frequent feedback ($d = 0.32$). Smither, London, and Reilly (2005) found that the degree of performance improvement after feedback appeared greater when reassessment came less than one year after the initial measurement than when it came one year or longer after the initial measurement.

Lam, DeRue, Karam, and Hollenbeck (2011) had participants complete a simulation task and found that there was no significant direct relationship between

feedback frequency as a predictor of the outcomes of task performance and task effort. However, they did find a significant curvilinear relationship between feedback frequency and the two dependent variables after controlling for participant gender, race, and feedback sign, such that initial increases in feedback frequency improved task performance and effort but "too much" feedback decreased performance and effort. It should be noted, however, that this investigation was conducted on a relatively small sample (N = 86). Kinicki, Prussia, Wu, and McKee-Ryan (2004) found a positive correlation between the frequency of feedback and job performance (r = 0.30).

Immediacy. Early researchers in feedback—specifically Knowledge of Results (KR) feedback—spent much time on examining the effect of the duration of delay between actual performance and the receipt of feedback on its effectiveness (Salmoni, Schmidt, and Walter, 1984). Results generally indicated that there was no effect of feedback delay on subsequent performance. This concept has been studied with great interest in education—Kulik and Kulik (1988) in their meta-analysis first noted that immediate feedback was more effective than delayed feedback when the feedback was given after each item (i.e., given immediately after the item was responded to versus with a few seconds delay after responding; d = 0.55) but when feedback was only given at the end of the test, immediate feedback was associated with decreased performance relative to delayed feedback (i.e., at test completion versus after a day). These results have been replicated a number of times (e.g., Fajfar, Campitelli, & Labolita, 2012). A parallel example in I-O psychology might be that feedback immediately after completion of a subordinate (i.e., low-level) goal might be more effective than delayed feedback, but delayed feedback after completion of a superordinate goal might be most effective.

Northcraft, Schmidt, and Ashford (2011) recently tested the transference of the effect of immediacy of feedback on relative resource allocation toward multiple goals. They found a main effect of immediacy on resource allocation such that participants worked more toward goals for which there was timely (i.e., more immediate) feedback. A main effect of immediacy was also found on task performance, such that performance on the task was higher under well-timed feedback conditions than in less timely feedback. They also found interaction effects between feedback specificity and feedback timeliness such that for vague feedback, immediacy had no effect; while timing had a significant positive relationship with performance and resource allocation under conditions of specific feedback.

Researchers have further investigated this phenomenon in field samples. Kuvaas, Buch, and Dysvik (2016) examined the relationships between perceived immediacy and frequency of supervisory performance feedback (assessed as one measure, i.e., quality), perceived constructiveness of the feedback, and subsequent job performance. They found evidence for an interaction effect of perceived feedback quality and perceived constructiveness on job performance, but no main effect of perceived feedback quality on job performance after controlling for gender, education, and tenure. The interaction was such that there was no relationship between constructiveness and work performance in individuals that perceived their feedback to be low in quality, but there was a positive relationship in individuals that perceived their feedback to be high in quality.

Specificity. Feedback specificity refers to the amount of information that is presented in feedback messages—it is related to the directive purpose of feedback (Bilodeau, 1966). Early research (e.g., Goldstein, Emanuel, & Howell, 1968; Johnson,

Perlow, & Pieper, 1993) found evidence that more specific feedback could result in increased job performance. This finding was further cemented by Kluger and DeNisi (1996)—feedback which provided the correct solution (i.e., that is more specific) was more effective at increasing job performance (d = 0.43) than feedback that did not provide the solution (d = 0.25). Davis, Carson, Ammeter, and Treadway (2005) found that high specificity feedback was much more highly correlated with performance than moderately specific feedback. Kinicki, Prussia, Wu, and McKee-Ryan (2004) found a significant positive relationship between feedback specificity and job performance (r = 0.24). Recent results from Northcraft, Schmidt, and Ashford (2011) found that resources tended to be allocated toward tasks with specific feedback available in addition to performance in those tasks tending to be higher.

Goodman and colleagues (Goodman & Wood, 2004; Goodman, Wood, & Hendricks, 2004) have examined the effects of feedback specificity during practice on both performance during the practice task and subsequent performance on a related task (i.e., learning). They have generally found that more specific feedback is positively related to performance on the task for which it is generated, but does not necessarily affect subsequent learning. Goodman and Wood (2004) found that specific feedback was effective at teaching good performers that what they were doing was correct (i.e., participants doing mostly the right thing increased their correct behavior on the learning task), but was not effective at teaching poor performers what they were doing wrong (i.e., participants doing mostly the wrong thing did not correct their behavior for the learning task).

Task Type. Kluger and DeNisi (1996) noted that feedback researchers had, for the most part, ignored the effect of the type of the task as a potential moderator of feedback's effectiveness. They hypothesized that the fewer the cognitive resources necessary for task performance (i.e., the simpler the task), the stronger the relationship between feedback and performance. Indeed, they found that task complexity was negatively related to the effect size (r = -0.11) but also found that feedback had much less of an effect on improving performance on physical tasks (d = -0.11) than on non-physical tasks (d = 0.36) and a greater effect on memory tasks (d = 0.69) than on non-memory tasks (d = 0.30). Similarly, Pritchard, Harrell, DiazGranados, and Guzman (2008) found that the ProMES system had the greatest effect sizes in studies of technical jobs (d = 2.15), followed by academic and managerial jobs (d = 1.74), and blue collar jobs (d = 1.54), and it was least effective in clerical jobs (d = 0.27). In contrast, Pritchard et al. also found that ProMES was least effective in manufacturing organizations (d = 1.05) and most effective in either sales (d = 1.45) or service organizations (d = 1.63).

Only one study directly examining the effects of task complexity on performance improvement published since the literature search by Kluger and DeNisi (1996) was found. Korsgaard and Diddams (1996) examined the interactive effects of task complexity and feedback availability on performance improvement. They found that for less complex tasks, the availability of feedback did not affect performance improvement. In the complex task condition, they found that participants only showed significant performance improvement if both process and outcome feedback[3] were available (compared to conditions where only outcome feedback was available). Van Dijk and

[3] Process and outcome feedback to be discussed in detail below

Kluger (2004; 2011) examined the moderating effects of the regulatory focus of tasks (i.e., prevention-focused versus promotion-focused[4]) on the relationship between feedback and performance improvement. They found no significant interaction effect of task type on feedback and performance improvement. However, they did find an interaction effect between task type and feedback sign on performance such that negative feedback increased performance for a prevention-focused task but decreased performance on a promotion-focused task. Inversely, positive feedback increased performance on a promotion-focused task but decreased performance on a prevention-focused task.

Medium. The method by which feedback is conveyed was studied with some regularity in the early days of feedback research. Balcazar, Hopkins, and Suarez (1986) found that there were more consistently positive effects in studies of graphic feedback (i.e., including illustrations; 54%) than there were in studies of written (32%) or verbal (24%) feedback. Alvero, Bucklin, and Austin (2001) found similar results—studies that combined either verbal or written feedback with graphs had more consistently positive effects than did methods without them. Kluger and DeNisi (1996) found that verbal feedback was less effective (d = 0.23) than non-verbal feedback (d = 0.37), but no significant effects for graphical or written feedback. They also compared computerized versus non-computerized feedback, finding that computerized was more effective (d = 0.41) than non-computerized feedback (d = 0.23). Later, Adler and Ambrose (2005) found that feedback given in a face-to-face setting (rather than through a computer) was

[4] A prevention-focused task is one that is framed by preventing a loss (e.g., working at a job that you have to keep), whereas a promotion-focused task is on that is framed by gains and accomplishments (e.g., working at a job that you desired to have).

positively associated with job performance (albeit indirectly, through perceived fairness of the feedback).

Privacy. FIT suggests that because it would threaten self-esteem, public display of feedback would decrease subsequent performance. However, research findings suggest otherwise. Balcazar, Hopkins, and Suarez (1985) found no real difference in the consistency of effects in studies of feedback for public, private, or some combination of the two. These findings were replicated in Alvero, Bucklin, and Austin (2001), with a small positive effect for feedback interventions using a mixture of public and private elements (however there were only 10 studies). Initial meta-analytic estimates from Kulger and DeNisi (1996) suggested that public feedback interventions had a greater effect than private (correlation with $d = 0.20$), but after exclusions of suspect studies, the effect was rendered non-significant (correlation with $d = 0.06$).

Mediators. Ilgen, Fisher, and Taylor (1979) supposed that there were a number of mediating processes between feedback and subsequent performance: perceptions of the feedback, acceptance of the feedback, desire to respond to the feedback, and an intention to respond to the feedback. They reviewed the early research on these variables, and recent investigations have further confirmed the existence of these mediating factors. For example, Jawahar (2010) found that the ratee's perceived accuracy of the feedback, perceived utility of the feedback, and the satisfaction with the feedback partially mediated the effect of feedback characteristics on job performance. Similarly, Anseel and Lievens (2009) found that feedback acceptance partially mediated the relationship between feedback and subsequent performance. Kinicki, Prussia, Wu, and McKee-Ryan (2004) explicitly tested the Ilgen, Fisher, and Taylor (1979) model of feedback in a

sample of bank employees and found evidence for positive, mediated relationships between perceived accuracy, desire to respond, intent to respond, and performance.

Measurement Considerations

The measurement of job performance as a criterion in selection has been a concern in Industrial/Organizational psychology since perhaps its inception (Blum & Naylor, 1968). Early work focused on finding the "ultimate criterion... the complete final goal of a particular type of selection or training" (Thorndike, 1949, p.121)—which most interpreted to mean it were possible to obtain one single measure that could wholly define job success in a given position. It was not until Dunnette (1963) told the field to "junk *the* criterion!" (emphasis in original, p. 252) in favor of building an understanding of the meaning of performance—with an emphasis on behavior—that researchers began to consider multiple criteria. Then, in the 1980s, researchers began to consider job performance as a multidimensional psychological construct deserving study of its own (Campbell, 2012). As a result of this work (e.g., Campbell, McCloy, Oppler, & Sager, 1993; Yukl, Gordon, & Taber, 2002), much attention has been given to its measurement. Campbell (2012) briefly described a number of measurement considerations in performance ratings which will be discussed in detail, along with other concerns not mentioned, in the sections below.

Instrument Concerns

In order to conduct a performance appraisal, there is typically some formal method or instrument of feedback collection. There are many considerations when designing a feedback instrument for performance appraisal, reviewed below. Specifically,

evidence for changes in rating behavior (e.g., leniency or severity) based on qualities of the measurement instrument will be reviewed.

Item and Scale Complexity. Viswesvaran, Ones, and Schmidt (1996) first noted the counterintuitive finding that for job performance ratings, an increase in the number of items did not seem to substantially increase intrarater or interrater reliability. They offered two potential explanations: (1) that the relationship between the number of items and reliability is convex such that after a scale reaches a certain length there are minimal gains in increased reliability or (2) that the broad nature of the job performance construct leads to more reliable ratings, as broad constructs had been shown to be more reliably rated than narrower ones (Ones & Viswesvaran, 1996). Not long thereafter, Greguras and Robie (1998) found similar results in their study of multisource feedback ratings—an increase in the number of items from three to twenty yielded minimal increases in reliability—as much as 0.09—when the number of raters was held constant. Wanous and Hudy (2001) later estimated the minimum reliability of a single-item, individual level performance measure at a reasonably high level (0.70).

Relatedly, the amount of inter-rater agreement in job performance ratings seems mostly unaffected by the number of items on the scale. Conway and Huffcutt (1997) found nearly identical reliabilities of job performance ratings made by supervisors, peers, and subordinates whether the scale was a composite or a rating of overall performance. Heidemeier and Moser (2009) replicated these findings in their meta-analysis of job performance ratings: agreement between self and supervisor ratings of performance was about the same for single-item measures of performance ($\rho = 0.34$) as it was for aggregated measures ($\rho = 0.32$).

A few researchers have investigated the complexity of the item itself as a source of rating variability. Brutus and Facteau (2003) examined the effects of item syntax (i.e., complexity), double-barreledness, linguistic specificity, and behavioral specificity on the psychometric quality (operationalized as the amount of variance in the item that was accounted for by the factor it was intended to measure) of multisource feedback items. They only found a negative effect of item syntax on item quality—more linguistically complex items were less related to the factor they were intended to measure. Kaiser and Craig (2005) examined the same variables (with the exception of linguistic specificity) but used interrater reliability and agreement as their outcome variables. They found no effect of any of the variables on interrater agreement, but did find that multi-barreledness and abstraction (i.e., low behavioral specificity) were negatively related to interrater reliability of peers and subordinates. In a related study, Roch, Paquin, and Littlejohn (2009) surprisingly found that items that were rated as less behaviorally observable (by upper-level PhD students) had higher levels of agreement than did more behaviorally observable items.

Topic. The content of the measurement instrument is also an important consideration to make. Annett (1969) first distinguished between feedback that concerned the outcomes of actions and feedback about the actions themselves. This distinction later became known as process feedback and outcome feedback (Earley, Northcraft, Lee, and Lituchy, 1990). Earley et al. (1990) examined the interactive effects of the type of feedback given and goal setting on task performance and found that the highest level of performance was associated with having both a specific, challenging goal as well as having access to both types of feedback. Also, as previously mentioned, Korsgaard and

Diddams (1996) they found that participants only showed significant performance improvement in a complex task if both process and outcome feedback were available (compared to conditions where only outcome feedback was available).

The dimensionality of the assessment also been considered. Harris and Schaubroeck (1988) found that there was slightly higher agreement between self- and supervisory-ratings of job performance for dimensional (ρ = 0.36) than for global ratings (ρ = 0.29). They found a moderator effect of dimensionality on agreement between peer-supervisor ratings, but in the opposite direction (ρ = 0.57 for dimensional and ρ = 0.65 for global ratings). No effect of dimensionality on the agreement between self- and peer-ratings of performance was found. Later, Heidemeier and Moser (2009) found that for single-item measures, self-other agreement was similar for global performance items (ρ = 0.33) as it was for specific performance dimension items (ρ = 0.34).

Other variables related to the content of the instrument have been examined, with somewhat less research interest. For example, the reviews by Balcazar et al. (1985) and Alvero et al. (2001) both examined the consistency in effectiveness of various combinations of feedback content (e.g., individual versus group performance, whether a standard of performance was included, and whether normative information was present) but found little differences. Harris and Schaubroeck (1988) also looked at the effect of trait-based versus behaviorally based ratings on self-other agreement and found that behaviorally-based scales had higher agreement for self-supervisor ratings (ρ = 0.43) than did trait-based scales (ρ = 0.32). For peer-supervisor agreement the opposite was found: there was more agreement in trait-based scales (ρ = 0.64) than in behaviorally-based scales (ρ = 0.53). Again, no interaction effect was found for self-peer agreement.

Response Scale. Researchers have been interested in the differences in ratings stemming from the response scale of the instrument for a long time (e.g., Taylor & Wherry, 1951). Landy and Farr (1980) noted that much of the research in the performance appraisal space at the time was conducted on the differential effectiveness of different rating formats. After a review of the research on graphic rating scales (GRS, Paterson, 1922), behaviorally anchored rating scales (BARS, Smith & Kendall, 1963), behavioral observation scales (BOS, Latham & Wexley, 1977), and forced choice (FC) scales, they conclude that "after more than 30 years of serious research, it seems that little progress has been made in developing an efficient and psychometrically sound alternative to the traditional graphic rating scale" (p. 89). They went on to say that only 4%-8% of the variance in performance ratings could be explained by the format—they even suggested a "moratorium on format-related research" (p. 101). Many researchers heeded this call. Others did not.

Research published near the time of Landy and Farr's (1980) review confirmed their assertion of relative equivalence of rating scales. Fay and Latham (1982) found no difference between BOS and BARS in terms of reducing rater errors. Murphy, Martin, and Garcia (1982) found that when performance was recalled after some delay, there was no difference in ratings made in BOS and in GRS. However, a later meta-analysis conducted by Jawahar and Williams (1997) found that the difference between ratings made for administrative purposes and those made for research purposes was the smallest for ratings with BARS (d = 0.03), followed by those made with FC scales (d = 0.18), and largest for ratings with GRS (d = 0.34). Other recent research has found that rater

personality had less of an effect on ratings made on behavioral checklists than it did on ratings made on GRS (Yun, Donahue, Dudley, & McFarland, 2005).

Other researchers have taken time to develop new rating scale formats in order to improve the psychometric quality of performance ratings. Borman, Buck, Hanson, Motowidlo, Stark, and Drasgow (2001) developed the Computerized Adaptive Rating Scale (CARS) format specifically to improve criterion measurement in I-O Psychology. In their example, they created a CARS measure of citizenship performance (i.e., Organizational Citizenship Behavior, OCB; Organ, 1988). The basic methodology is borrowed from computerized adaptive testing (CAT; Weiss, 2011). In this case, item parameters were generated by subject matter experts rather than by algorithm. For each of the three OCB dimensions, up to 15 pairs of behavioral statements were shown to respondents, and they were asked to select which was more accurate in describing the ratee. In terms of inter-rater reliability, the CARS format ($ICC_{2,1}$ = 0.78) was marginally larger than it was for the BARS ($ICC_{2,1}$ = 0.74) and GRS ($ICC_{2,1}$ = 0.73) formats. However, validity (i.e., the accuracy of the raters' rank order) was considerably higher in the CARS condition (r = 0.84) than it was in either BARS (r = 0.49) or GRS (r = 0.72) conditions. They also found that all four of Cronbach's (1955) measures of rating accuracy[5] were better in the CARS condition than in the other measurement format conditions. Schneider, Goff, Anderson, and Borman (2003) expanded the CARS methodology to measure managerial performance, but unfortunately did not empirically test the format against others.

[5] Described in detail in the "Rater Concerns" section

Kaplan and Norton (1992, 1996) developed the balanced scorecard (BSC) format. Originally, it was created to give top-level managers a quick and comprehensive view of their organization's performance. Basically, a BSC provides an individual with organizational goals and performance metrics in four areas: finance, internal business, innovation and learning, and customer perspectives. Researchers later developed BSCs at the individual level and examined how these systems might influence the eventual performance appraisal of subordinates (e.g., Banker, Chang, & Pizzini, 2004; Kaplan, Peterson, & Samuels, 2007). However, the applicability of this system outside of specific job domains (e.g., accounting) has limited the research potential of this rating format.

Very few studies have examined the role of narrative comments in the feedback process. Smither and Walker (2004) examined the role of the number, focus, and favorability of comments in a multisource feedback intervention on performance improvement. They found that individuals improved the most under conditions of a low number of comments that were unfavorable, that were about behaviors or tasks rather than about traits. In contrast, individuals with a high number of unfavorable comments about behaviors or tasks tended to decrease in performance. David (2013) found similar results in a sample of nurses—receipt of favorable comments and comments with high levels of interactional justice (i.e., treating the employee with dignity, respect, kindness, and consideration) was positively related to a measure of performance collected one year later. However, she found no effect for comment length, specificity, or whether it contained content specific to a performance goal. Somewhat tangentially, Wilson (2010) content-analyzed the comments that accompanied numerical ratings of a sample of hotel staff employees to examine the differences in supervisory comments across ratee ethnic

groups. She found that supervisors tended to give positive comments, regardless of the numerical rating that they gave. That is, there was no substantial difference in the content of comments made to higher performers and those given to low performers. She also found that mention of interpersonal and social factors of performance were more likely to occur in appraisals of minority (i.e., Black or Asian) employees (David, 2013).

The response scale of the performance appraisal instrument has also been linked to goal-setting behavior. Tziner and Kopelman (1988) found that ratings made on a BOS yielded higher levels of goal acceptance, goal commitment, and goal clarity than did ratings made on a GRS. Tziner, Kopelman, and Joanis (1997) found that the clarity of the path to the goal, goal acceptance, goal commitment, and goal specificity were all rated higher under ratings given in BOS scales than those in either BARS or GRS. In contrast, Tziner, Joanis, and Murphy (2000) found that although ratees were more satisfied with ratings made with BOS than with BARS, ratings made on a GRS were as good as BOS and as good or better than BARS in terms of goal specificity, goal observability, and ratee's perceptions of goals.

Medium. The method by which the feedback is communicated has been suggested to play a part in the accuracy of performance ratings. Specifically, some have hypothesized that feedback that is delivered face-to-face may be more challenging for a rater than delivery via a non-face-to-face vehicle. Indeed, Klimoski and Inks (1990) found that individuals that expected to deliver their feedback anonymously were considerably more strict (i.e., more accurate in their ratings) than individuals that expected to deliver their feedback face-to-face. Waung and Highhouse (1997) also showed that when participants expected to give their feedback in a face-to-face setting,

they were more likely to give positive feedback than when they expected to give feedback indirectly (through a video-taped recording, in this instance).

In contrast, Smither, Walker, and Yap (2004), in a study of upward feedback, looked for a potential elevation effect when managers selected that their subordinates give them feedback via the company's intranet compared to "traditional" paper-and-pencil feedback. Initial results suggested that ratings were more favorable for online than paper-pencil versions (d = 0.38), but after accounting for rater and ratee characteristics (e.g., age, tenure), this effect was diminished. Similarly, Yun, Donahue, Dudley, and McFarland (2005) found no direct effect of medium (face-to-face, identified versus aggregated, anonymous feedback) on rating elevation. However, a three-way interaction between performance level of the ratee, agreeableness of the rater, and medium was found. Raters of low performers that were high on agreeableness elevated their scores more when they were not expecting to give face-to-face feedback; those that were low on agreeableness elevated their ratings more when they were expecting to give face-to-face feedback. For raters of moderate performers, the effect was reversed. Raters of high performers that were high on agreeableness elevated their ratings when they expected to give face-to-face feedback—those that were low on agreeableness were unaffected by the feedback medium.

Other researchers have examined more distal variables in relation to feedback medium. Adler (2007) examined the effects of feedback that was provided by computer or provided by a supervisor on the perceived interpersonal fairness of the feedback. He found a non-significant direct effect of source on fairness perceptions, but did find an interaction effect of constructiveness (i.e., inconsiderate, general versus specific, non-

threatening) and source on fairness—under conditions of constructive feedback, getting feedback from the supervisor was perceived as more fair than getting it from the computer. There was no difference in perceived feedback fairness in the destructive feedback condition. Au and Chan (2013) compared the relative preferences of employees in Hong Kong for using face-to-face, telephone, email, or written communication methods to provide feedback. They found that individuals that were low in communication orientation were more likely to use email or written feedback; face-to face and written feedback was more likely to be used when communicating with subordinates and the phone was more likely to be used with peers; positive feedback was more likely to be conveyed via phone, email, or via written communication.

Contextual Concerns.

Performance appraisals do not happen in a vacuum—they are conducted within the context of not only the immediate relationship between manager and supervisor, but also within that of the team and organization. Landy and Farr (1980), in their review of the performance-appraisal literature, mention that context is an important component in performance appraisal, but also note that the amount of research in the space is limited. Furthermore, Murphy and Cleveland (1995) observed that: "the effects of context variables on appraisal processes and outcomes have been the object of speculation but have not been empirically examined in the detail that these effects warrant" (p. 407). Their call to action seems to have been effective—Levy and Williams (2004) and Ferris, Munyon, Basik, and Buckley (2008) were able to procure sufficient research on the topic to conduct substantive literature reviews on context effects in performance appraisal, and

research in this space has continued. An overview of their findings, and a review of the literature since that time, will continue below.

Purpose. Meyer, Kay, and French (1965) noted the inherent multiplicity in purpose of performance appraisal in applied settings—they are designed to both provide justification for administrative decisions (e.g., salary increases, promotions) and simultaneously motivate the employee to improve their performance. They noted that often "the traditional appraisal system essentially becomes a salary discussion in which the manager justifies the action taken" (Meyer, Kay, & French, 1965, p. 129). Cleveland, Murphy, and Williams (1989) provided some data to further illustrate this point. Responses from a survey of *Society of Industrial Organizational Psychology* (*SIOP*) members suggested that most organizations frequently—and simultaneously—use performance appraisals for four distinct purposes: (1) to compare between individuals on their performance level, (2) to identify potential strengths and weaknesses within an individual, (3) to implement and evaluate human resource systems and interventions, and (4) to document or justify personnel decisions.

Taylor and Wherry (1951) were among the first to hypothesize that performance appraisal ratings made for administrative purposes would be more lenient than ratings for other purposes. A meta-analysis of the literature confirmed this suspicion: Jawahar and Williams (1997) found strong empirical support for an effect of purpose on leniency (mean unweighted $d = 0.25$) such that administrative ratings were substantially higher than research or developmental ratings. Ratings that were made in the field (i.e., by actual managers of actual employees) were considerably inflated (mean unweighted $d = 0.32$). These results have also been replicated in a laboratory setting (Curtis, Harvey, & Ravden,

2005) Harahi, Rudolph, and Laginess (2015) further found that personality of the rater generally has a greater influence on the level of rating given in administrative contexts than it does in research and developmental contexts.

Relatedly, Heidemeier and Moser (2009) found that agreement between self-ratings and supervisory ratings was highest in administrative conditions ($\rho = 0.42$), followed by research settings ($\rho = 0.35$), then in conditions of explicitly combined purposes ($\rho = 0.32$), then in those made for developmental purposes ($\rho = 0.24$). Finally, Smither, London, and Reilly (2005) demonstrated that feedback given in a developmental context was more effective at improving performance ($d = 0.25$) than that given in an administrative context ($d = 0.08$)—regardless of the source of the feedback (i.e., supervisor, peer, or direct report). Most recently, Ellington and Wilson (2016) found that the level of supervisory ratings was considerably higher in administrative settings than in developmental settings in a police department.

Prior Information. Murphy, Balzer, Lockhart, and Eisenman (1985) were among the first to consider the potential for contrast and assimilation effects in the performance appraisal process. Information about past performance was hypothesized to impact ratings of present performance in one of two ways: (1) present evaluations are biased in an opposite direction to previous performance (a contrast effect) or (2) present evaluations are biased toward the direction of previous performance (an assimilation effect). In their study, they found strong contrast effects for previous performance ratings—individuals that "improved" in their performance were rated higher than individuals that "declined" in performance. However, after they delayed the observation of previous and present performance, the effect went away. Mero and Motowidlo (1995)

found that when raters were told that the person they were rating had been rated "too low" in the past rated their subordinates more favorably—demonstrating evidence for a contrast effect. However, Salvemini, Reilly, and Smither (1993) found evidence for an assimilation effect in prior performance ratings.

Smither, Reilly, & Buda (1988) expanded their method to include the means of procuring prior information about the performer as a variable of interest. They found evidence for a contrast effect when knowledge of prior performance was directly observed, but an assimilation effect when prior performance was communicated via pervious performance ratings. However, when there was more of a delay between viewing of initial (good or bad) performance and later (average) performance, there were no effects. Similarly, Huber, Neale, and Northcraft (1987) found no difference in the ratings made by managers in either the presence or absence of prior performance relative to performance standards (e.g., "in the past, this person performed above average").

Maurer, Palmer, and Lisnov (1995) examined the effects of another individual's performance on the rating of performance (e.g., the effect of a peer's performance level on an evaluation of another). They demonstrated that the context in which performance appraisals are made can influence rating behavior; when average performance of one person was preceded by good performance of another, raters were more severe. Similarly, when poor performance of an individual was followed by average performance of another, raters were much more lenient.

Thorsteinson, Breier, Atwell, Hamilton, and Privette (2008) have demonstrated that assimilation effects can occur with information that is not at all relevant to the current measurement of performance. They showed three separate samples of participants

only one example of a filled-out performance rating instrument that was either rated at a high level (i.e., all ratings were 9 out of 9) or at a low level (i.e., all ratings were 1 out of 9) of performance before asking them to complete a performance appraisal for a lecturer's performance. Raters that were anchored by high levels of performance (made after seeing the example with high performance ratings) gave higher ratings than control group raters (who saw no example) and low-anchored raters (who saw the example with low performance ratings).

Other researchers have examined the effects of behavioral expectations of the rater on the level of rating. For example, Mount and Thompson (1987) found that ratings are more accurate when the behaviors of the ratee are congruent with the behavioral expectations of the rater, however, they also found greater leniency and halo under these conditions. Hogan (1987) found a significant direct effect of supervisory expectations on predicting performance six months later—those that were expected to do well were rated highly. In addition, she found an interaction effect between actual and predicted performance on performance ratings such that those with low expected performance but actual high performance were rated lower than those that were predicted to have high performance. Similarly, those that were expected to have high performance, but had low actual performance were actually rated lower than those who were met expectations of low performance. Thus, there is evidence to suggest that managers tend to provide ratings that are in-line with their expectations.

Self-Appraisal Information. Some organizations choose to have their employees provide self-ratings of performance in addition to their manager's ratings. Research has shown that if self-appraisal information is made available to a rater, it can

affect their ratings. Early results from Klimoski and Inks (1990) suggested that raters tended to gravitate toward the self-rating of performance as given by the performer. Shore and Tashchain (2002) examined the effects of the presence of self-appraisal information, normative information, and actual performance on ratings. They found a main effect for both task performance and self-assessment: higher task performance was associated with higher ratings, as were higher self-assessments. They also found interaction effects of task performance and self appraisal. Regardless of self-appraisal level, high task performance was rated equally. However, when a poor performer provided a high self-rating, the rater provided a higher rating of performance than if the performer provided a low (i.e., more accurate) self-rating. Randall and Sharples (2012) replicated these results—raters gave higher ratings to poor performers when the performers gave themselves inaccurately high self-ratings than when they gave themselves moderately inaccurate and accurate self-ratings of performance.

Normative Information. Some performance measurement instruments encourage the rater to consider a comparison group (e.g., "consider the average employee") or sometimes provide an explicit standard of performance (e.g., "an average performer behaves in this way"). Shore and Tashchain (2002) found an interaction effect of the presence of normative information and the level of self-rating on the rater's given performance rating: when a norm (i.e., standard) of performance was included, ratings were equal regardless of self-appraisal level. However, when no norm was provided, individuals that gave themselves higher self-appraisals scored higher than individuals that gave themselves low self-appraisals. In contrast, in the meta-analysis by Heidemeier and Moser (2009) no difference in self-other agreement was found when raters were told to

consider a comparison group or were given a standard (ρ = 0.36) and when raters were not given a standard (ρ = 0.32).

Politics. Longenecker, Sims, and Gioia (1987) brought the issue of the role of politics in performance appraisal to light. They interviewed a number of executives across a number of organizations, and came to four conclusions: (1) that politics is a reality of organizational life (and performance appraisal is no exception); (2) that many things influence the political culture of an organization (e.g., its economic health, executive beliefs about performance management); (3) that performance rating inflation happens for many reasons, some of which are political (e.g., to promote "up and out" a person with bad fit to avoid confrontation); and (4) that performance rating deflation happens for many reasons, some of which are political (e.g., to send a message to the subordinate to leave, to "teach a subordinate a lesson"). Gioia and Longenecker (1994) interviewed more executives and derived five similar themes: (1) that the higher one rises in the organization, the more political the performance appraisal process becomes, (2) appraisals of managers are particularly susceptible to politicking, (3) raters are affected by things other than performance when making the rating (e.g., maintaining a reputation, the political climate of the organization), (4) senior executives have too much freedom in their evaluations, and (5) appraisal is a political tool to control people and resources.

In order to encourage quantitative research in this area, Tziner, Latham, Price, and Haccoun (1996) developed a questionnaire (the Questionnaire of Political Considerations in Performance Appraisal, QPCPA; later renamed the PCPAQ; Tziner, Prince, & Murphy, 1997) for measuring the influence of political considerations in performance appraisal. This instrument was developed specifically to be completed by raters of

performance in order to understand the political climate under which raters operate, and demonstrated decent test-retest reliability and both convergent (with Machiavellianism and Need for Power) and divergent (with organizational commitment) validities. Tziner (1999) then used the measure to understand what leads to use of political considerations in performance evaluations: he found that self-efficacy as a rater and continuance commitment were both negatively related to reliance on politics in a police officer sample.

Bernardin and Beatty (1984) did not study politics per se, but did examine the effects of rater perceived norms regarding the rating process on rating level. They developed the "Trust in the Appraisal Process Survey" (TAPS) which assessed the degree to which a rater thinks that a "typical supervisor" intentionally manipulates their performance ratings of their subordinates. In a quasi-experiment in two police departments, they found that, over time (i.e., six months after the PA system was in place), there was less trust in the accuracy of the appraisals as well as a general increase in the level of the performance ratings, and that these problems were exacerbated in the department where appraisals were used for administrative purposes compared to the department where the ratings were used only for feedback purposes. A similar study was recently conducted by Spence and Keeping (2010). They found that when raters suggested that the organizational norms (as perceived by the raters) were to provide high ratings, the ratings were indeed higher than when organizational norms were to provide accurate ratings. Also, when giving high ratings was in the rater's best interest (i.e., when their subordinates' ratings were a component of their performance rating), ratings were higher. Unfortunately, few other researchers have investigated the effects of politics from

this perspective—most have examined the effects of the ratee's perceptions of organizational politics (POP; Ferris, Russ, & Fandt, 1989).

While the ratee's perspective on organizational politics (i.e., the factors they perceive are affected by politics in the organization) has long been demonstrated to predict important outcomes (e.g., job satisfaction, turnover, and organizational commitment; see Miller, Rutherford, & Kolodinsky, 2008 for a recent meta-analysis), it is only relatively recently that researchers have investigated a link between POP and performance ratings. Researchers have repeatedly found a direct negative correlation between a ratee's perception of organizational politics and job performance as rated by their supervisor—individuals in a more political environment tend to have lower performance (Chen & Fang, 2008; Witt, 1998; Vigoda, 2000; Zivnuska et al., 2004). They tend to suggest that the effect is due to a highly political climate *causing* decreased performance, usually through posited low expectations of ratings being influenced by actual performance. However, this research is entirely cross-sectional, within one organization, at one point in time—it is also possible that low performers tend to "explain away" their low performance by indicating that their environment is dictated by politics and not by merit. Therefore, while it seems logical that politics play a part in performance ratings, their exact mechanisms are not well understood.

Accountability. Schlenker, Britt, Pennington, Murphy, and Doherty (1994) outlined the necessary conditions for an individual to feel responsible for an action: (1) the action should have a set of governing instructions or rules of conduct, (2) the person is bound to behave within the confines of those rules, and (3) the person has control over the action. Accountability, then, "involves an evaluative reckoning [of these three

elements] in which the individuals are judged" (p. 634)—it requires an audience. Over the years, researchers have hypothesized the effects of accountability in both directions—Bernardin, Thomason, Buckly, and Kane (2016) hypothesized that high accountability would result in a higher influence of personality on rating inflation, whereas Harahi, Rudolph, and Laginess (2015) suggested that high accountability would lead to less personality-influenced rating inflation. Empirical results have also been in either direction. For example, Klimoski and Inks (1990) found that ratings were less accurate when participants were required to justify their responses, whereas Mero and Motowidlo (1995) found the opposite. Similarly, Bernardin et al. (2016) found evidence for a stronger effect of personality under highly accountable conditions in their study, whereas Harahi et al. (2015) found the opposite in their meta-analysis.

Results from Curtis, Harvey, and Ravden (2005) can elucidate the problem. They noted the difference between upward and downward accountability: upward accountability is when the rater is held responsible by someone of higher standing than themselves (e.g., the researcher, or their supervisor) whereas downward accountability is when the rater is held responsible by a party of lower standing than themselves (e.g., their subordinate). Klimoski and Inks (1990) and Bernardin et al. (2016) held their participants accountable to their ratees, whereas Mero and Motowidlo (1995) held them accountable to the researchers. Harahi et al. (2015) did not provide a clear description of their accountability definition, but it is possible that they used the various authors' definitions that went into their meta-analysis, as the variance around their estimate is very high. In experiments that explicitly defined (and manipulated) accountability along the upward-downward dimension found that upward accountability resulted in more accurate ratings,

while downward accountability resulted in more lenient ratings (Curtis, Harvey, & Ravden, 2005; Mero, Guidice, & Brownlee, 2007). One incongruent result does exist, however: Spence and Keeping (2010) found that lower ratings were given when the manager was expecting to be confronted by the ratee. They posited that the expectation of a confrontation by the employee (in a vignette) may have led to disliking of the employee[6].

Ratee Concerns

Certain characteristics of the ratee might influence the rating that they receive from their supervisor. Some are relatively stable individual differences; others are characteristics of the job itself. Ideally, the only influential characteristic in a performance rating would be the ratee's actual level of job performance, but there is some research to suggest that that is not the case. Evidence will be reviewed regarding the influence of individual differences (e.g., demographics, personality), the ratee's job type, and ratee ability.

Individual Differences. There are certain stable characteristics of the ratee that may influence the ratings that are given to them by their rater. Some of the most researched variables are demographic variables (e.g., race, gender, and age). This is most likely due, in part, to passage of the Civil Rights Act of 1964—specifically Title VII—which prohibits discrimination in employment decisions like hiring and "status as an employee" on the basis of race, color, religion, sex, and national origin. Age was later protected under the Age Discrimination in Employment Act of 1967. Other individual difference variables (e.g., personality) have also been studied.

[6] "liking" effects to be discussed below

Demographics. Early work on the effects of demographics on received ratings was mostly conducted in the laboratory on undergraduate student participants that read vignettes or watched recorded videos of performance. Some found a significant effect of sex on the performance ratings they received: Hamner, Kim, Baird, and Bigoness (1974) found that female ratees were rated more highly than were males, particularly at high levels of performance (Bigoness, 1976 replicated these results); Rosen and Jerdee (1974) found that females were rated much lower than were males; Hall and Hall (1976) found no effect of the gender of the ratee on performance ratings. Similarly, Hall and Hall (1976) found no effect of race on the performance ratings received, but Bigoness (1976) found that low-performing blacks were rated higher than low performing whites. Similarly, Rosen and Jerdee (1976a, 1976b) found that older employees were judged as less cognitively, physically, and emotionally fit to perform.

Eventually researchers moved out of the laboratory and began to use field samples. Ferris, Yates, Gilmore, and Rowland (1985) found that older nurses were rated much lower than their younger counterparts. In a sample of sales representatives, Liden, Stilwell, and Ferris (1996) found that older employees were rated more highly than younger employees. Both Waldman and Avolio (1991) and Landau (1995) found that, after controlling for certain demographic variables (e.g., age, education, tenure), female employees were rated lower on promotion potential than were men; similarly, black and Asian employees were rated lower than were white employees. Greenhaus, Parasuraman, and Wormley (1990) found that black employees were rated lower on both job performance and promotability. In contrast, Pulakos, Schmitt, and Chan (1996) found no significant difference in their measurement model fit between different races or genders,

and Bass and Turner (1973) found no significant difference between ratings of bank tellers of different races. This mixture of findings led some researchers to search for the answer with two approaches: large sample studies and meta-analyses.

Pulakos, White, Oppler, and Borman (1989) found significant main effects of ratee race on performance rating received in a large military sample: for technical skill and job effort, Hispanics were rated the highest, followed by whites, followed by black personnel; for personal discipline, Hispanics and whites were rated the highest, followed by blacks; for military bearing ratings, blacks and Hispanics were rated more highly than white employees. McKay and McDaniel (2006) later conducted a meta-analysis on black-white differences in performance ratings and found evidence of higher ratings for whites (d = 0.27), and the difference was larger for overall ratings of job performance (d = 0.35) than for other measures (e.g., task performance, d = 0.21). In their meta-analysis, Bowen, Swim, and Jacobs (2000) found little evidence for gender bias in performance ratings in operational settings (d = -0.01). Roth, Purvis, and Bobko (2012) updated this analysis and included promotability ratings—females were generally rated higher on job performance (d = 0.11), but men were rated higher on promotability ratings (d = 0.10).

Job Tenure. It seems logical that the longer an individual is on a job, the better they would perform that job, and therefore the ratings that a more experienced worker gets may be larger than those of a less experienced worker. Early results from Zedeck and Baker (1972) confirmed this suspicion: they found a significant correlation between tenure as a nurse and performance ratings (r = 0.23). However, many more results have suggested that there is no effect of job tenure of the ratee on the performance rating that they are given: Bass and Turner (1973) found positive relationships between time on the

job for bank tellers and supervisory ratings of certain dimensions of performance (e.g., cooperation, $r = 0.25$) but not for the overall performance rating ($r = -0.06$, n.s.). Cascio and Valenzi, (1977) found no effect of ratee experience on BARS scales of police officer performance. Huber, Neale, and Northcraft (1987) found no significant correlation between ratee job experience and performance rating for a mixed group of ratee jobs (e.g., administrative, technical, and professional positions). More recently, Greguras (2005) demonstrated measurement equivalence of a performance rating scale across different levels of ratee managerial experience, and also found that the mean level of ratings (of adaptability and coaching skills) was similar across all managerial experience levels.

Personality. The personality of the ratee has garnered some attention in predicting supervisory ratings of performance, particularly in the ratings of expatriate employees. Expatriate performance was hypothesized to be influenced by personality factors like open-mindedness, extraversion, and emotional stability due to the nature of expatriate life (Arthur & Bennett, 1995; Mendenhall & Oddou, 1985; Hammer, Grudykunst, & Wiseman, 1978). However, in an explicit test of the Big Five personality factors on expatriate supervisory performance ratings, Caligiuri (2000) found that only ratee conscientiousness was positively related to supervisory performance ratings in an expatriate assignment.

Others have examined the effects of personality on ratings of job performance for non-expatriate assignments. In a military sample, Borman, White, and Dorsey (1995) found that factors like dependability and friendliness were not associated with performance ratings. Tett, Jackson, and Rothstein (1991) conducted a meta-analysis on

the use of personality as a predictor of job performance (specifically to speak for its use as a predictor in selection) and found that there was generally a modest correlation between the Big Five personality factors and job performance ratings—the effects were smaller in incumbent samples than they were in recruited samples, but were also larger in studies of subjective criteria (i.e., ratings) than they were in studies with objective job performance criteria. Barrick and Mount (1991) found similar results in their meta-analysis. Judge, LePine, and Rich (2006) found no relationship between the Big Five personality factors and peer ratings of leadership or supervisory ratings of either contextual or task performance. They did find a significant negative effect of ratee narcissism on peer ratings of performance, but no direct effects on either type of supervisory rating. Bell and Arthur (2008) found that assessor ratings in an assessment center were not related to the ratee's extraversion (r = 0.15, n.s.), emotional stability (r = 0.12, n.s.), or agreeableness (r = 0.01).

Job Type. The nature of the ratee's job also has some influence over the measurement of performance in that job. Several meta-analyses have borne out this relationship. The meta-analysis of Harris and Schaubroeck (1988) showed that across-source correlations were generally higher for blue-collar or service jobs than they were for managerial or professional jobs, particularly for self-supervisor correlations (ρ = 0.27 in managerial/professional, ρ = 0.42 in blue-collar/service jobs). Similarly, Conway and Huffcutt (1997) found higher interrater reliabilities for supervisory ratings of performance and peer ratings of performance when jobs were lower in complexity (e.g., ρ = 0.60 in low complexity jobs versus ρ = 0.48 in high complexity jobs) and in non-managerial jobs (ρ = 0.54 versus ρ = 0.44 in managerial jobs). Heidemeier and Moser

(2007) found similar results in correlational agreement between the self and supervisor: blue-collar jobs had higher agreement (r = 0.33) than did white-collar jobs (r = 0.21); less complex jobs had much higher agreement (r = 0.47) than did complex jobs (r = 0.29).

Ability. The link between cognitive ability (CA, a.k.a. general mental ability, GMA) and supervisory ratings of job performance has been established empirically numerous times. Hunter (1983) conducted a path analysis which demonstrated that GMA is positively associated with job knowledge, which is in turn positively associated with supervisory ratings of performance. Schmidt, Hunter, and Outerbridge (1986) replicated these results in both a civilian and a military sample. Coward and Sackett (1990) provided empirical evidence that the relationship between cognitive ability and job performance ("typically supervisor ratings", p. 298) was linear and positive. Hunter and Schmidt (1996) and Schmidt and Hunter (2004) provided even more evidence to suggest that ability and job performance are positively related. Other groups of researchers (e.g., Kolz, McFarland, & Silverman, 1998; Morgeson, Delaney, & Hemingway, 2005) have also confirmed the relationship. It is perhaps one of the least controversial statements in I-O Psychology that ability predicts job performance (which is most frequently operationalized as supervisory ratings).

Rater Concerns

There are many qualities of the raters themselves which can heavily influence the quality of the performance rating that they give. Cronbach (1955) developed a method of assessing rating accuracy that mathematically separated the accuracy space into four measures: elevation, differential elevation, stereotype accuracy, and differential accuracy. Elevation is the extent to which a rater under- or over-rates a ratee compared to some

expert's rating. Differential elevation is the extent to which the rater can distinguish how far the ratee deviates from the average ratee's performance. Stereotype accuracy describes the rater's ability to determine what the average performance is. Differential accuracy is the rater's ability to rate the differences between ratees on a given item. Although this method is no longer used with much frequency, it exemplifies the need for accurate ratings in research. An understanding of what affects a rater's accuracy is essential to building knowledge in the feedback research space.

Individual Differences. Research has demonstrated that rating elevation (i.e., leniency) is a relatively stable characteristic (Borman & Hallam, 1991; Kane, Bernardin, Villanova, & Peyrefitte, 1995; Villanova, Bernardin, Dahmus, & Sims, 1993). As such, a number of relatively stable individual differences have been investigated in an attempt to understand the potential mechanisms of rating elevation's stability.

Personality. Bernardin and colleagues (Bernardin, Cooke, Villanova, 2000; Bernardin, Tyler, & Villanova, 2009; Kane et al., 1995; Villanova et al., 1993; Villanova, Bernardin, & Ross, 1997) have studied the effects of rater personality on rating effects for several years using student samples. Villanova, Bernardin, and Ross (1997) used the 16 personality factor model (16PF; Cattell, Eber, & Tatsuoka, 1970) to assess peer-rating leniency levels—their results suggested that agreeableness was positively related to rating level, but conscientiousness was not. Bernardin, Cooke, and Villanova (2000) repeated the study, but used the Big Five Factor model (FFM; Costa & McCrae, 1992). They found that (1) agreeableness was positively correlated with rating level, (2) conscientiousness was negatively correlated with rating level, and (3) raters that were both low on conscientiousness and high on agreeableness had ratings that were more

elevated than all other groups combined. Bernardin, Tyler, and Villanova (2009) replicated these results and additionally found that agreeableness was negatively associated with and conscientiousness was positively related to rating accuracy. Most recently, Bernardin, Thomason, Buckley, and Kane (2016) found an inverted U-shaped relationship between rating accuracy and both agreeableness and assertiveness.

Yun, Donahue, Dudley, and McFarland (2005) increased the fidelity of their study by training their student sample (with frame-of-reference training) on how to rate performance on their ratee's task, and also by including a condition where raters were expected to give the feedback to the ratee in a face-to-face meeting. They found that agreeableness was positively correlated with rating elevation, particularly in the face-to-face feedback condition. Dewberry, Davies-Muir, and Newell (2013) further increased fidelity by using trained assessors in medical assessments; unlike most performance appraisal situations, however, the raters were not expected to interact with the ratees after the appraisal process was complete. They found no effect of personality on rater leniency or severity.

Ogunfowora, Bourdage, and Lee (2010) used a policy capturing design to understand how rater personality affects the relative weights that raters put on aspects of performance when making a singular rating. They found that raters high on Openness to Experience placed greater weight on adaptive performance, while raters high on Modesty (they used the HEXACO model of personality; Ashton & Lee, 2007) weighted deviant behaviors more heavily than those low on the dimension. Bono, Hooper, and Yoon (2012) conducted two studies on this topic: one field observational study of leaders and one experimental study of students. In their first study, they found a positive relationship

between four of the five personality factors (agreeableness, openness, extraversion, and conscientiousness) and rating level, but no relationship between neuroticism and rating level. In their experiment, they found no effect for any of the personality traits.

Harari, Rudolph, and Laginess (2015) conducted a meta-analysis on 28 studies of the relationship between the Big Five personality traits and performance ratings. Results from bivariate analyses suggest that agreeableness, extraversion, and emotional stability were all positively related to the level of performance rating (ρ = 0.25, ρ = 0.12, and ρ = 0.14, respectively). Results for conscientiousness were highly variable (80% CV [-0.12, 0.32]). Multivariate regression analyses predicting performance rating behavior found significant results for agreeableness (β = 0.22), extraversion (β = 0.08), emotional stability (β = 0.08), and openness (β = -0.06).

Demographics. Some early research in the performance appraisal literature focused on the effect of gender of the rater on ratings. Many researchers found no effect of rater gender on performance ratings (Huber, Neale, & Northcraft, 1987; Mai-Dalton, Feldman-Summers, & Mitchell, 1979; Mobley, 1980; Nieva & Guteck, 1981; Rosen & Jerdee, 1976b), but typically used small student samples. Although they too used a student sample, Hamner, Kim, Baird, and Bigoness (1974) found that females gave higher ratings than men did. These findings were later replicated by Mobley (1982), Wexley and Pulakos (1982), and Pulakos, White, Oppler, and Borman (1989) also in a supply organization sample, a mixed organization sample, and an Army personnel sample, respectively. However, Northcraft, Huber, and Neale (1988) found the opposite to be true: in their sample female supervisors gave lower ratings than male supervisors. Others (Shore & Thornton, 1986) found no effect of gender on performance ratings.

There are large sample-sized and methodologically sound studies on both sides of the effect; researchers later turned to studying relational demography (discussed in the section Rater-Ratee Relationship. below) before settling this debate.

The research on the effects of rater age is more scant than for rater gender but is more conclusive. Early results from Klores (1996) suggest that age has no effect: immediately preceding its inclusion in a list of independent variables it was not "discussed further because, except for the expected obvious correlation between age and years with the company, there were no indications of any significant or even consistent relationships between age and experience variables and performance ratings" (414). Later, Rosen and Jerdee (1976a) observed that for one of four performance dimensions, older raters perceived a smaller difference between young and old ratees than did younger raters. Similarly, Schwab and Heneman (1978) found that there was a significant effect of rater age for only one out of six types of performance ratings such that older raters gave higher ratings than younger raters did. Later, in a sample of managers, Cleveland and Landy (1981) found that *younger* supervisors in a manufacturing organization rated their subordinates more highly on one performance dimension (Interpersonal Skills), and rater age had no effect on the other five dimensions. Finkelstein and Burke (1998) failed to find any connection between rater age and performance ratings, but did find an effect for ratings of economic worth. The lack of consistency in the findings, and the fact that often there is an effect for only one of many dimensions, suggests that there is likely no effect of rater age on performance ratings.

The effect of the race of the rater also garnered some attention around the same time as rater gender, but was usually included as a supplementary analysis to a gender

inquiry. Hamner, Kim, Baird, and Bigoness (1974) found that black raters tended to give higher ratings than did white raters. However, their sample was rather small and was conducted in a laboratory setting (i.e., they were non-consequential ratings). Schmitt and Lappin (1980) examined both the level of the performance rating and the accuracy of the rating. They found no effect for the race of the rater in terms of the average level of the performance rating between black and white raters, but did find that there was a higher correlation between rated performance and actual performance in the white rater group than in the black rater group. Both studies were based on relatively small student samples that rated pre-recorded task performance by actors. In contrast, Mobley (1982) examined actual performance ratings from supervisors in a supply organization and found no effect of rater race on the ratings that they gave.

Experience. Although job (or organizational) tenure is typically highly correlated with age, it has been examined a number of times as a separate predictor of rating behavior. Early results from Klores (1966) suggest that there was no significant effect of rater tenure with the company or tenure with the ratee group on the ratings given. Cascio and Valenzi (1977) had highly experienced (six or more years in-grade) and less experienced (less than six years in-grade) metropolitan police sergeants rate their subordinate officers on their performance. They found that more experienced sergeants gave higher performance ratings than did their less-experienced counterparts. Huber, Neale, and Northcraft (1987) found no effect of job tenure on performance ratings. Similarly, Govaerts, van de Wiel, and van der Vleuten (2013) found no effect of rater experience on the quality of feedback that was given.

Other researchers have examined the effect of amount of experience with performance appraisals on rating behavior. Spence and Keeping (2010) found that individuals that had more experience with performance appraisal, in general, gave lower ratings than those with less experience. Villanova, Bernardin, Dahmus, and Sims (1993) developed the Performance Appraisal Discomfort Scale (PADS) and demonstrated that rating leniency was highly correlated to responses on the scale—those that were less comfortable giving performance ratings were more likely to be lenient. However, Tziner, Murphy, Cleveland, Yavo, and Hayoon (2008) found no significant relationship between scores on the PADS and absolute rating level or ability to discriminate between ratees. Bernardin, Thomason, Buckley, and Kane (2016) designed the "Employee Discussion" exercise—in which a manager roleplays an interview with a store associate and is observed by two assessors—to evaluate Performance Management Competence (PMC). They found that PMC was negatively related to mean rating level given by the manager and was positively associated with rating accuracy.

Ability. Rater cognitive abilities have been studied extensively as a correlate of rater accuracy. Early research results suggested that there was a positive linear relationship between rater intelligence and rating accuracy (Borman, 1979). Smither and Reilly (1987) found that rater intelligence was significantly positively correlated with stereotype accuracy (i.e., rater ability to determine the average). They also found a curvilinear relationship between rater intelligence and differential accuracy (i.e., rater ability to determine ratee's deviation from average on an item) and stereotype accuracy such that those that were particularly high and particularly low on intelligence were less accurate than those that were moderately intelligent. Later researchers were able to

replicate the general finding of a linear relationship between rater intelligence and accuracy, but not the curvilinear relationship (Borman & Hallam, 1991; Hauenstein & Alexander, 1991).

The aforementioned studies had many characteristics in common: they were all conducted in laboratory settings with vignettes or videos of relatively objective task performance (e.g., jet engine installation) and intelligence was measured with the Wesman Personnel Classification Test (WPCT; Wesman, 1965). In contrast, Bartels and Doverspike (1997) found that intelligence of an assessment center rater (as measured by the 16PF; Cattell, Eber, & Tatsuoka, 1970) was positively related to rating leniency. Dewberry, Davies-Muir, and Newell (2013) found no correlation between fluid intelligence (as measured by a version of Raven's Progressive Matrices) and rating leniency or severity of clinical examination assessors.

Commitment. Tziner, Murphy, Cleveland, Beaudin, and Marchand (1998) hypothesized that organizational commitment level of the rater would increase the probability that raters would invest more effort into the performance appraisal process and therefore be positively related to discrimination among ratees and performance dimensions, and be related to a willingness to assign low scores. They examined this relationship in a relatively small student sample (N = 121) and found that organizational commitment was not significantly related to rating level nor related to discrimination among ratees. They were encouraged by the relatively high magnitude of the correlations, however, and obtained a second sample in which to test the relationship again. Tziner and Murphy (1999) found a significant positive relationship between commitment and ability to discriminate between ratees, but not between commitment and rating level or

discrimination among dimensions. However, in a set of seven samples using structural equation modeling, Tziner et al. (2001) found that the combined effect of organizational commitment and climate perceptions were positively related to rating level and discrimination among dimensions, but was negatively related to discrimination among ratees.

Training. The potential need for rater training was recognized very early in the history of I-O psychology (Bitner, 1948). Early reviewers of the literature suggested that rater training is effective at reducing rater bias and at increasing rating accuracy (Smith, 1986; Spool, 1978). Woehr and Huffcutt (1994) expanded their review to assess the effectiveness of rater training across the types of training and dependent variables. They defined four basic approaches to rater training: (1) rater error training, which familiarizes raters with the basic types of psychometric errors (e.g., leniency, halo, central tendency) and encourages them to avoid committing such errors; (2) performance dimension training, which acquaints raters with the various dimensions of performance on which employees should be rated so as to avoid making global, heuristic-based ratings; (3) frame-of-reference training, which is typically performance dimension training but also includes behavioral examples of each level of performance in each dimension so that raters are using a common conceptualization of performance when making their ratings; and (4) behavioral observation training, which combats lack of observation on the part of the rater—raters are trained on their ability to recall behavioral events and are assessed for their recall accuracy. The dependent variables assessed in the meta-analysis were: (1) halo error, (2) leniency error, (3) rating accuracy (e.g., the distance between the trainee's

an expert's rating), and (4) observational accuracy (e.g., the ability to correctly recall the number of times a behavior occurred).

Their results indicate that rater error training is the most popular form of training: it was effective at reducing both halo and leniency errors (d = 0.33 and d = 0.21, respectively) and increased rater accuracy (d = 0.26). Interestingly, they found that rater error training which trains raters to recognize rating patterns as potentially erroneous (i.e., with a focus on correcting their distribution of ratings) actually *decreased* rating accuracy (d = -0.20) while traditional rater error training without the emphasis of rating distributions substantially increased rating accuracy (d = 0.76). Performance dimension training had a positive effect on decreasing halo error (d = 0.30), but a slight increase on leniency (d = -0.14) and a small increase in rating accuracy (d = 0.13). Frame-of-reference training was the most effective at increasing rating accuracy (d = 0.83), with slight decreases in both halo (d = 0.13) and leniency (d = 0.15) errors. Behavioral observation training was effective at increasing rating and observational accuracy (d = 0.77 and d = 0.49, respectively), but had very few data points available (k = 4, N = 224).

Motivation. The importance of understanding the motivation of a rater was recognized by Taft (1955):

> But probably the most important area of all is that of *motivation*: if the judge is motivated to make accurate judgements about his subject and if he feels himself free to be objective, then he has a good chance of achieving his aim. (p. 21, emphasis original).

However, in the performance appraisal domain, the consideration of the motivational aspects of the process was not a topic of interest until considerably later (cf. Landy &

Farr, 1980). Bernardin and Beatty (1984) believed that the lack of evidence for the effect of rating scale formats on rating accuracy was due largely to the fact that the "largest portion of the variance in format comparisons lies with individual raters and their motivation (or lack thereof) to rate accurately" (p. 267-268). They did not provide any data to support this supposition, however.

Harris (1994) provided a motivational framework to guide research into the role of rater motivation on performance appraisal behaviors—situational factors (e.g., accountability, trust) and personal factors (e.g., self-efficacy and mood) he posited would influence the motivational factors (rewards, negative consequences, and impression management) of performance appraisal, which would, in turn, affect the performance appraisal behaviors (observation, storage, retrieval, integration, rating, and feedback). Although the model was somewhat simple, it seemed to help spur research interest into the area—a decade later, there was enough research to justify a review.

Spence and Keeping (2011) first noted that there were many names to describe similar motivations for performance appraisal behavior; they condensed the somewhat separate research areas of politics, impression management, leniency, rater goals, and rater motivation down to a few basic motivations. Raters were found to inflate ratings: (1) to avoid confrontation with subordinates, (2) to comply with organizational norms, (3) to promote a problem employee out of the department, (4) to appear as a competent manager, (5) to procure resources, (6) to motivate or act in the best interest of the ratee, and (7) because performance appraisal competes with other, more rewarded tasks. Raters had also been found to deflate ratings, mostly to build a case to justify firing of an

employee or to "send a message" to poor performers. Soon after, Spence and Keeping refined their theory.

Spence and Keeping (2013) built out their findings to fit within the context of the theory of planned behavior (TPB, Ajzen, 1991). They suggested that there were four rater intentions that could influence their subsequent ratings: to avoid conflict, to be benevolent, to be accurate, and to impression manage. The relative weighting of these intentions is determined by the rater's attitudes toward the intention, their perceived norms for the intention, and the perceived behavioral control over the intention. The attitudes, norms, and control of the rater are influenced by so-called "tributary variables" such as manager characteristics, subordinate characteristics, situational factors, and subordinate job performance. An outline of their model is shown in Figure 3.

Other Characteristics. A number of other rater characteristics have been studied with somewhat less frequency. For example, the mood that the rater is in has been shown to affect the ratings that they give—generally, raters in positive moods tend to give higher ratings than raters in negative moods (Fried, Levi, Ben-David, Tiegs, & Avital, 2000; Poon, 2001). Further, Robbins and DeNisi (1998) found that raters in a negative mood recalled more performance incidents, and that raters in general were more likely to recall events that were incongruent with their current mood (e.g., a rater in a negative mood was more likely to recall positive performance behaviors from their subordinate).

Other researchers have examined the effect of rater philosophical perspectives on the ratings that they give. Wexley and Youtz (1985) measured rater beliefs about human nature (e.g., human trustworthiness, independence, rationality, altruism, and variability), and found that raters that believe that people are basically good and that people are more

or less the same are the most lenient, while those who believe that people greatly differ from one another rated the most accurately. Heslin and colleagues (Heslin, Latham, & VandeWalle, 2005; Heslin & VandeWalle, 2008) have also studied the effects of rater perceived malleability of personal attributes (i.e., the extent to which a rater believes that a person's personality and abilities are changeable) on rating behavior. They found that raters who perceive that personal attributes are less malleable are less likely to recognize actual changes in employee performance and are less likely to coach employees on how to improve their performance.

Rater-Ratee Relationship. To this point, simple main effects of rater and ratee characteristics on the ratings that are given or received have been discussed. However, interaction effects between rater and ratee characteristics, and also characteristics of the relationship itself have also been investigated. The most popular of these will be discussed below: rater liking of the ratee, rater-ratee similarity, and the duration of the rater-ratee relationship.

Liking. A great deal of research has been conducted on the effects of interpersonal affect or "liking" of the ratee—that is, the extent to which the rater's liking of the ratee can influence their ratings. Many researchers have found evidence for a positive relationship between rater interpersonal affect and rating level (e.g., Cardy & Dobbins, 1986; Turban, Jones, & Rozelle, 1990; Conway, 1988; Judge & Ferris, 1993; Tsui & Barry, 1986), while some found no effect (e.g., Borman, White, & Dorsey, 1995). Others searched for moderators—for example, in-person performance versus electronic performance (Weisband & Atwater, 1999), feedback source (Antonioni & Park, 2001); rating content (Varma, DeNisi, & Peters, 1996), and country culture (Carmona, Iyer, &

Reckers, 2014). Some investigated explanatory variables like likelihood to punish poor performance (Dobbins & Russell, 1986) or recollection of positive and negative performance (Robbins & DeNisi, 1998).

In order to better understand the relationship between interpersonal affect and performance ratings, Sutton, Baldwin, Wood, and Hoffman (2013) conducted a meta-analysis—they found a substantial positive relationship between liking and performance ratings (ρ = 0.77). They also found significant moderator effects for the performance construct that was measured, the source of the ratings, the purpose of the ratings, the job complexity of the ratee, and whether the ratee's job was in sales or not. Measures of OCB were considerably less correlated with liking (ρ = 0.51) than were measures of task performance (ρ = 0.75) or of overall performance (ρ = 0.77). Peer ratings were the least correlated with liking (ρ = 0.63), followed by supervisory ratings (ρ = 0.72), and subordinate ratings were the most correlated (ρ = 0.72). Ratings that were used for developmental purposes only were least correlated with liking (ρ = 0.60), followed by those for research purposes (ρ = 0.70), then by ratings for administrative purposes (ρ =0.71). Ratings made for ratees in less complex jobs were less correlated with liking (ρ = 0.59) than they were for ratees in more complex jobs (ρ = 0.66). Interestingly, ratings were more associated with liking in ratings of sales jobs (ρ = 0.65) than they were for non-sales jobs (ρ = 0.54). It also seems that in spite of these significant moderator analyses, there is a considerable association between liking and performance ratings.

However, there is limited research into whether this association between rater interpersonal affect and performance ratings is a source of bias or if it reflects an association between liking and actual performance (i.e., raters tend to like good

performers more than they like poor performers). Lefkowitz (2000) noted this issue in his qualitative review, and Sutton et al. (2013) also acknowledged it in their discussion. They both suggest that the existing research tends to find that actual performance and liking are highly related, but that most of it is conducted within laboratory or controlled settings and used a variety of methodologies that make strong conclusions difficult (Adams, 2005; Allen & Rush, 1998; Robbins & DeNisi, 1994; Varma, Pichler, & Srinivas, 2005; Varma & Pichler, 2007; Wayne & Ferris, 1990).

Similarity. Tsui and O'Reilly (1989) coined the term "relational demography" to refer to the relative similarity between the demographic categories (e.g., race, age, education, tenure) to which a supervisor and subordinate belong. They examined the effects of relational demography in the aforementioned variables on job performance ratings and rater interpersonal affect—after controlling for subordinate and supervisory demographics alone, whether the dyad being of the same gender and of similar job tenure significantly predicted performance ratings. For affect, relational demography was significantly predictive in gender, education, and job tenure. Ratees in mixed-gender and mixed-job tenure dyads were rated lower and less liked than ratees in matched dyads. Similarly, those from mixed educational backgrounds were less liked by their supervisor than those in matched educational backgrounds. Some later results suggest that there is no effect of relational demography of age on supervisory performance ratings (Vecchio, 1993; Zalesny & Kirsch, 1989), other results suggest that there may be an effect, through increased liking (Judge & Ferris, 1993). Educational similarity has been shown to increase job performance ratings (Zalesny & Kirsch, 1989), but also has received null evidence (Cascio & Valenzi, 1977). Interestingly, Strauss, Barrick, and Connerley (2001)

found that actual relational demography on personality had no effect on performance ratings, but perceived personality similarity did. Also, Carmona, Iyer, and Reckers (2014) found that perceived similarity to a ratee was more influential in a sample from Spain than it was in a sample from the United States.

The demographic difference in race of rater-ratee dyads is among the more widely studied topics in this area. An early meta-analysis by Kraiger and Ford (1985) was particularly influential—they found that white raters tended to rate white ratees more highly than they did black ratees ($d = 0.37$) and black raters rated black ratees more highly than they did white ratees ($d = 0.45$) there was, however, considerable variability in the effect sizes. In addition, they found a significant moderator effect of race of the ratee for white raters in lab settings ($d = 0.07$) than in field settings ($d = 0.39$). They did not, however, find significant moderator effects of rater training (present or not), rating scale (behavioral versus trait), or for rating purpose (administrative versus research).

However, as Pulakos, White, Oppler, and Borman (1989) note, Kraiger and Ford (1985) perfectly confounded race of the rater and of the ratee—that is, ratees were only rated by a rater of one race; it could be argued that ratees that happen to work for a black rater are lower performers. To address this, they used a large military sample in which ratees were rated by two raters—one of each race. They found a significant interaction effect of ratee race and rater race, although their effect sizes were considerably smaller (e.g., $r = 0.08$ for white supervisors, where a positive correlation indicates favorability toward white ratees, $r = 0.02$ for black supervisors, same direction) than were Kraiger and Ford's. Thus, there was some evidence to suggest that white ratees tend to be rated higher, regardless of the race of the rater.

Sackett and DuBois (1991) acquired data that were similar to Pulakos et al. (1989), but in a civilian sample. They found a similar result—white raters tended to rate white ratees more highly than they rated black ratees (d = 0.36), but black raters also rated white ratees more highly than they rated black ratees (d = 0.16) suggesting that individuals do not, in fact, rate subordinates more highly if they are from the same race. They further discredit the Kraiger and Ford (1985) results by pointing out that many of their meta-analyzed studies were from laboratory, undergraduate rater samples, and that half of them came from peer ratings of performance, rather than supervisory ratings. Some later results from Geddes and Konrad (2003) suggest that same-race dyads are not always better: in fact, employees reacted more negatively to negative feedback from same-race managers—but they generally reacted more positively to feedback from white managers.

Duration. Researchers, particularly those in the study of Leader-Member Exchange (LMX, Graen & Uhl-Bien, 1995) have considered the length of the relationship as a potential influence on performance ratings. Early results suggested that longer rater-ratee relationships were associated with higher rating levels (Duarte, Goodson, & Klich, 1994; Vecchio, 1998). Smith, Harrington, and Houghton (2000) hypothesized that the effect may be due, in part, to increased comfort with giving feedback with time; however, the length of relationship did not affect performance appraisal discomfort. Similarly, Vanneste, Puranam, and Kretschmer (2014) found that the amount of trust in the supervisory relationship and its duration is positively, but modestly correlated (ρ = 0.12) and highly variable across studies. Rothstein (1990) found a non-linear relationship between the months of exposure to the ratee and interrater reliability of their

supervisors—for ratings of ability, interrater reliability asymptotes at about 0.60, and reaches that level after about a year.

Interrater Reliability. Another consideration in the measurement of performance is that of interrater reliability. The value of including multiple raters in the assessment of job performance was recognized early in the history of I-O psychology (Lawler, 1967)—it can provide evidence of convergent and discriminant validities, as well as build understanding into how the appraisal process works.

A brief distinction between interrater reliability and interrater agreement ought to be made before proceeding. The extent to which raters agree with one another has been operationalized in a number of ways—many of them which actually reflect interrater reliability. Interrater reliability is best assessed by correlational measures (e.g., Pearson's *r*, intraclass correlations)—it provides an indication of the extent to which the rankings made by a number of raters are the same (Kozlowski & Hattrup, 1992). Interrater agreement, on the other hand, is an assessment of the fungibility of raters—the extent to which one rater can be replaced with another and the absolute value of the rating does not change. Admittedly, the waters here are quite muddy, particularly when it comes to operationalization of the two variables in the literature as it currently stands, and the topic does not go without its own controversy (see Kozlowski and Hattrup, 1992).

Within Sources. Interrater reliability within sources refers to the relative accord between raters in similar relative positions to the ratee (e.g., reliability among supervisors). Viswesvaran, Ones, and Schmidt (1996) conducted the most thorough meta-analysis on interrater reliability in performance ratings. They estimated the supervisory interrater reliability for overall job performance ratings was moderate ($\rho = 0.52$) and that

ratings of particular performance dimensions were, at times, higher (e.g., Quality ratings, $\rho = 0.63$) while other times lower (e.g., Communication competence, $\rho = 0.45$) but not considerably so. Peer ratings of performance were, in general, estimated to be less reliable than those of supervisors ($\rho = 0.42$), with the exception of ratings of compliance ($\rho = 0.71$, peers; $\rho = 0.56$, supervisors). Ratings of overall job performance by supervisors were also estimated to be highly stable ($\rho = 0.81$). Conway and Huffcutt (1997) meta-analyzed multisource performance rating reliabilities and found similar results to Viswesvaran et al. (1996). They also expanded to include subordinate ratings—subordinates were the least reliable ($\rho = 0.30$), followed by peers ($\rho = 0.37$), and supervisors ($\rho = 0.50$). Greguras and Robie (1998) also conducted a meta-analysis in this space—their estimates were quite similar for supervisory ratings ($\rho = 0.51$), for peer ratings ($\rho = 0.37$), and for subordinate ratings ($\rho = 0.35$).

Between Sources. Reliability between sources refers to the correlation across multiple groups of individuals (e.g., reliability between supervisors and peers). Early meta-analytic results from Harris and Schaubroeck (1988) suggest that peer-supervisor ratings have the highest correlation ($\rho = 0.62$), followed by self-peer ratings ($\rho = 0.36$) and self-supervisor ratings ($\rho = 0.35$). Then, Conway and Huffcutt (1997) found that the correlations between the self and other groups were the lowest ($\rho = 0.19$ with peers, $\rho = 0.22$ with supervisors, and $\rho = 0.14$ with subordinates), the subordinate ratings and other groups were next ($\rho = 0.22$ with peers, $\rho = 0.22$ with supervisors), and the supervisor-peer ratings were the most highly correlated ($\rho = 0.34$). Most recently, Heidemeier & Moser (2009) conducted a meta-analysis of self and supervisory ratings of job performance

outside of multisource ratings and found an estimate that was more similar to the results of Harris and Schaubroeck (1988; $\rho = 0.34$).

Intra-Rater Reliability. Intra-rater reliability refers to the estimated reliability of one rater from a given category (i.e., the rate-rerate reliability of a single rater that has been "Spearman-Browned" down to the reliability of one rating). Viswesvaran, Ones, and Schmidt (1996) found that intra-rater reliably were considerably higher than interrater reliabilities from the same group—supervisory overall job performance ratings and peer ratings were very high ($\rho = 0.86$ and $\rho = 0.85$, respectively). These estimates are much higher than those obtained by Greguras and Robie (1998; e.g., $\rho = 0.35$ for 1 supervisor) but that is likely due to the fact that in contrast to Viswesvaran et al. (1996), they used reliabilities from between two separate raters at one point in time, rather than one rater at two points in time.

Self-Other Agreement

The agreement between self-ratings of performance and other-ratings of performance is one area of the interrater agreement research space that has garnered much attention, particularly in the 360° feedback on managerial performance literature. So-called self-other rating agreement (SOA) has been operationalized in many ways but is usually defined as the degree of agreement or congruence between an employee's self-ratings and the ratings of others (e.g., in 360° feedback, ratings of the individual given by supervisors, peers, subordinates, clients, and customers; Yammarino & Atwater, 1993). Incongruence between self and other ratings can be, and has been, interpreted in a number of ways: (1) as an indicator of measurement error in performance ratings, (2) as an additional source of valid information regarding performance, and, more recently, (3)

as a marker of an individual's self-awareness. The latter two interpretations have become increasingly accepted as evidence accumulates that demonstrates that SOA is related to certain individual characteristics (e.g., personality; Fletcher & Baldry, 2000) as well as important outcomes of interest (e.g., performance and compensation; Ostroff et al., 2004).

As previously discussed, interrater reliabilities that include self-ratings tend to be the lowest (Harris & Schaubroeck, 1988). Self-ratings also tend to be the least valid when compared to objective criteria (Ashford, 1989; Mabe & West, 1982; Yammarino & Atwater, 1993). However, when Mabe and West (1982) conducted a meta-analytic moderator analysis, they found that individuals that were high in intelligence, high in achievement status, and had an internal locus of control were able to provide more accurate self-evaluations. This analysis was tangential to the variables that they were interested in, and therefore, went mostly undiscussed as a theoretical contribution. However, this result was merely the first step in building the models of self-other agreement.

Models of Self-Other Agreement.

Ashford (1989) summarized the then-relatively sparse literature on self-assessments in organizations and provided a comprehensive model to direct future research in the area. Researchers had focused on documenting discrepancies in self- and other-assessments of performance, and Ashford wanted her theory to inspire researchers to explain why discrepancies existed. An adaptation of her model is shown in Figure 4. She outlined three major tasks an employee has to complete in a self-assessment as well as three problems that might influence their ability to complete those tasks. Her three so-

called antecedent problems are: the information problem, where the amount and quality of information available to the employee can vary widely and be ambiguous; the ego defense problem, where an individual's self-esteem preservation may lead to avoidance of important feedback; and the self-presentation problem, where an employee's image of themselves may interfere with their ability to objectively evaluate feedback. The tasks that an individual must complete in a self-assessment are to: 1) establish the link between their own behavior and the standard of performance, 2) establish the link between their own behavior and the feedback cue, and 3) interpret the relevant cues correctly.

One outcome in her model of self-assessment is the agreement between the self and some other assessor—a 2-by-2 matrix of the agreement space: the self and the other determine whether the employee is "on track" or "off track" regarding their goal (See Figure 4). Individuals that are in agreement about their performance, Ashford suggested, would maintain their current strategies ("on track" agreement) or, in the event of "off track" agreement, alter their behavior to try to become "on track" or potentially abandon their goal. Individuals with inflated self-views were theorized to persist in their efforts until confronted with the others' views. Individuals with lower self- than other-assessments, on the other hand, may alter behavior that was actually effective in favor of behavior that may not be effective. Ashford also suggested that the organization may be affected by self-other discrepancies, but did not make any clear assertions as to their nature.

Yammarino and Atwater (1993) built upon Ashford's theory, focusing on the outcomes of self and other agreement. In their model, individuals could be categorized as: an over-estimator (i.e., they rate themselves more highly than others rate them), an

accurate estimator (i.e., they rate themselves at a similar level to how others rate them), or an under-estimator (i.e., they rate themselves lower than how others rate them). On the basis of previous results, they posited that over-estimation would lead to diminished organizational and individual outcomes, accurate estimation would lead to enhanced outcomes, and under-estimation would lead to mixed outcomes. Specifically, they suggested that agreement would affect: self-diagnoses of strengths and weaknesses, aspiration level, feedback acceptance and use, job satisfaction, attitudes toward supervisors, and goal attainment. Organizational outcomes included career derailment, promotions, leadership performance, and training and job performance. They also discussed potential individual (e.g., personality, ability, biodata) and environmental (e.g., the social environment, job factors) characteristics that might influence the likelihood of achieving self-other agreement.

Atwater and Yammarino (1997) developed their 1993 model further to include elements of Ashford's (1989) model (i.e., inclusion of self-perception accuracy and its determinants) as well as an expansion of the categorization of agreement to a four-group model. The model is summarized in Figure 5. The model begins with the accuracy of the perception of the self and its proposed determinants: biographical characteristics like age and tenure, individual characteristics like personality and locus of control, and cognitive processes like beliefs and schemas. This accuracy in self-perception then influences the accuracy in the actual rating of the self (i.e., an accurate self-rating is only possible with accurate self-perception). In addition, certain job relevant experiences (e.g., previous success or feedback) and contextual factors (e.g., political influences, comparative information) may influence the self-rating accuracy but not that of the self-perception.

These same determinants also affect the accuracy of the other's rating of the individual along with other factors like interpersonal biases (e.g., familiarity, similarity).

From there, individuals can be categorized as over-estimators, under-estimators, or in-agreement good estimators (i.e., agreement that their performance is good), or in-agreement poor estimators (i.e., agreement that their performance is lacking; similar to Ashford's (1989) model). They then suggested, in addition to the hypotheses from the 1993 model, that in-agreement good estimators would experience positive individual and organizational outcomes and in-agreement poor estimators would experience negative outcomes, but less negative outcomes than over-estimators. Outcomes were the same as those discussed in 1993, but expanded to include goal-setting behavior, absenteeism, organizational commitment, workplace conflicts, and subsequent job performance improvement. Their distinction between individual-level and organization-level outcomes is somewhat dated—they refer to variables like turnover, commitment, and leader performance as organizational outcomes, when they are typically measured at the individual level, and have been measured at this level for all of the investigations into this area to date.

Fleenor, McCauley, and Brutus (1996) expanded the four-group model to a six-group model: they basically expanded the good/poor distinction to both over-estimators and under-estimators (rather than only in the in-agreement group as in Atwater & Yammarino, 1997). However, the proposal of this model was more a methodological contribution than a theoretical one—there were no real propositions as to what might be expected or any evidence to support these claims. They simply categorized their sample of manager self-ratings and subordinate ratings (of the manager) into both the four-group

and six-group models and found that the conclusion that would have been drawn using the four-group model was not replicated in the six-group model. Modern methods of analysis (e.g., polynomial regression, discussed below) have mostly done away with the need to explicitly categorize individuals before analysis can be done, but these theories have aided in the interpretation of results and the building of hypotheses—much of the terminology remains the same, but its operationalization has changed dramatically.

Methods of Self-Other Agreement.

The methods of measuring self-other agreement have grown simultaneously with the building of its theories. Very early methodology relied on difference scores between self-ratings and other ratings. After this approach was repeatedly criticized by Edwards (1993, 1994, & 2002), its popularity substantially decreased. To take its place was the categorization method first employed by Atwater & Yammarino (1992) and later expanded upon by themselves and other researchers (Yammarino & Atwater, 1997; Fleenor et al., 2006). The basic approach is this: (1) obtain the average other rating for each manager, (2) transform self-ratings and other-ratings into z-scores within rater groups, (3) any person with a difference between the scores that exceeded a magnitude of 1 was considered over- or under-estimators (e.g., individuals that rated themselves more than 1 point higher than their other raters did were deemed over-estimators). The good/poor categorization was generally determined by the sign of the average other-rating. However, this approach had many limitations: it suffered from many of the same issues as difference scores, the exact method of categorization was never standardized, categories were sample-dependent (i.e., an individual may be considered an over-

estimator in one group but perhaps in-agreement in another), and the categorization needlessly made a continuous variable into a discrete variable.

Interestingly, Edwards suggested that either polynomial regression (1993, 1994, & 2002) or multivariate regression (1995) be used when evaluating self-other agreement, but the published research did not seem to incorporate these methods as standard until the mid-2000s. Multivariate regression is used when agreement is the outcome of interest and proceeds as any other multivariate regression. Polynomial regression is used when SOA is used as a predictor variable—a three-dimensional surface is constructed which allows researchers to keep the ratings as separate continuous variables while also considering their effects on the outcome of interest. The procedure has three steps: (1) regress the outcome variable on the additive function between the self-ratings and the other ratings, (2) add a term for squared self-rating, squared other-rating, and the interaction between self- and other-rating; if a significant change in R^2 is found, proceed to (3) conduct response surface tests. Typically, researchers examine both the slope and curvature of the lines of perfect agreement and of perfect disagreement to determine how agreement, over-estimation, and under-estimation of self-ratings relate to an outcome of interest.

Predictors of Self-Other Agreement.

There have been many investigations into which factors predict the likelihood of congruence between one's self-ratings and the ratings of others. The findings can be grouped into four categories: demographic characteristics, personality characteristics, contextual factors, and measurement characteristics. Findings in each of these areas will be discussed.

Demographic Characteristics. Gender of the target individual has been examined as a correlate of self-other agreement with some regularity. Male participants have been found to over-rate themselves more frequently than female participants in their leadership effectiveness (Brutus, Fleenor, & McCauley, 1999; Vecchio & Anderson, 2009), sales and marketing skills (Lindeman, Sundvik, & Rouhianen, 1995), job performance (Patiar & Mia, 2008), specific abilities (i.e., mathematical, spatial, and kinesthetic abilities; Visser, Ashton, & Vernon, 2008), and competency ratings (Jones & Fletcher, 2002). This finding is not entirely consistent, however. Some researchers have found that gender does not affect self-other agreement (e.g., Van Velson, Taylor, & Leslie, 1993) and others' data suggest that the effect is not due to males over-rating themselves but in fact is due to others rating females more highly than males (Ostroff, Atwater, & Feinberg, 2004). Another possible explanation is that women tend to be more accepting of and responsive to feedback from others (Roberts & Hoeksema, 1989; Roberts, 1991). Roberts and Nolen-Hoeksema (1994) found that this effect was due to women tending to view others' ratings of their performance as more accurate than men do, thus incorporating more information into their self-perception.

The age of the target individual tends to affect the degree of self-other agreement. Researchers have repeatedly found that older raters tend to over-rate themselves in many domains: in their job performance (Brutus, Fleenor, & McCauley, 1999; Ferris, Yates, Gilmore, & Rowland, 1985; Lawler, 1967), their scholastic achievement (Bailey & Bailey, 1974), their managerial effectiveness (Vecchio & Anderson, 2009), their transformational leadership (Moshavi, Brown, & Dodd, 2003), and their leadership behaviors (Ostroff, et al., 2004). However, Ostroff, Atwater, and Feinberg (2004) noted

that this effect in their data was due to older managers rating themselves more highly than did younger managers while the other raters (subordinates, peers, and supervisors) tended to rate older managers lower than they rated younger managers. Shore and Blieken (1991), on the other hand, found that middle-aged workers had the least self-supervisor agreement compared to older and younger employees.

The organizational level at which the person operates (e.g., middle manager, division lead, executive) is highly correlated with their age (Ostroff et al., 2004) and as a result has been found to be correlated with a lack of self-other agreement (Brief, Aldag, & Van Sell, 1977; Brutus, Fleenor, & McCauley, 1999; Gentry, Hannum, Ekelund, & de Jong, 2007; Ostroff et al., 2004; Sala, 2003). This effect may be due, in part, to decreased feedback-seeking behavior in more tenured employees compared to their less tenured counterparts (Ashford, 1986). Interestingly, however, leaders have been found to have higher levels of agreement with others' ratings than individuals in non-leadership roles (Gallo & McClintock, 1962; Green, 1948; Greer, Galanter, & Nordie, 1954; Lansing, 1957). Similarly, Bailey and Fletcher (2002) found that, over time, leaders tend to increase their agreement with their supervisors' and subordinates' ratings.

Only two studies have examined the race of the individual as a predictor of self-other agreement. Ostroff et al. (2004) found that non-white managers tended to rate themselves more highly than did white managers on their managerial effectiveness, but that others did not rate differentially based on race—suggesting that non-white managers tended to over-rate themselves more than white managers. However, Vecchio and Anderson (2009) found no effect of race on self-other agreement of managerial effectiveness. Both samples were large (>1,000), from multiple organizations, and used a

large-item inventory. However, Ostroff et al. (2004) aggregated others' ratings within categories whereas Vecchio and Anderson (2009) randomly selected one response from each category.

Finally, Ostroff et al. (2004) also examined the effects of education level and years of experience as a manager on self-other agreement. They found a positive correlation between years of experience as a manager and tendency to over-rate. Interestingly, they found a positive correlation between manager education level and level of ratings by themselves and by others (i.e., self- and other-rated performance trends upward with education level) and self-other agreement (i.e., more highly educated managers are more likely to rate themselves at similar levels as their other raters). Less educated managers tended to over-rate themselves.

Personality Characteristics. Only a few studies have examined the effects of personality traits on self-other agreement directly. Fletcher and Baldry (2000) found that certain sub-dimensions of the big five (as measured by the 16PF, Cattell, Eber, & Tatsuoka, 1970) were correlated with self-supervisor agreement—agreeableness (detached—outgoing) and openness (conservative—experimenting) were positively correlated and neuroticism (forthright—shrewd) was negatively correlated with self-other agreement. Brutus et al. (1999) found that only social dominance was negatively correlated with self-other agreement in all categories of other ratings, while other characteristics like social presence and communality were predictive of only self-peer or self-other agreement. Goffin and Anderson (2007) found that managers with higher self-esteem were more likely to over-rate their job performance.

Other researchers have examined the differential effect of personality on self and other ratings separately, but did not examine self-other congruence as an outcome. Self-other agreement was not the primary focus in Judge, LePine, and Rich (2006), but the results suggest that narcissism, openness to experience, agreeableness, and conscientiousness are all positively related to self-ratings of leadership behavior, but were not predictive of others' ratings. Similarly, Bell and Arthur (2008) found that extraversion was positively related to participants' self-ratings of performance but were not related to their assessors' ratings. Thus, there is some evidence to suggest that personality may be related to self-other agreement but few have directly examined these relationships.

Contextual Factors. The setting in which the ratings are being made can have a large influence on self-other agreement, generally through the deflation of self-ratings. If the self-rater is made aware that their ratings can or will be checked or compared to objective criterion measures, they tend to make more accurate ratings (Farh & Werbel, 1986; Heidemeier & Moser, 2009; Mabe & West, 1982; Regan, Gesselink, Hubsch, & Ulsh, 1975). Early results suggested that self-ratings tend to be inflated when used for evaluative (rather than developmental) purposes (Farh & Werbel, 1986), but some recent results indicate that ratings for developmental purposes tend to be inflated relative to ratings for research-only purposes (Heidemeier & Moser, 2009).

Similarly, culture of the individual (i.e., not organizational culture) has been shown to affect self-other agreement through differences in self-ratings. Individuals living in collectivist cultures tend to rate themselves lower than individuals from more individualistic cultures on leadership behavior (Atwater, Wang, Smither, & Fleenor,

2009) and on job performance (Farh, Dobbins, & Cheng, 1991; Farh & Cheng, 1997; Yik, Bond, & Paulhus, 1998). Similar results were found by Xie, Roy, and Chen (2006) when they measured individual orientations toward individualism/collectivism rather than using the country-level variables. In direct investigations of self-other agreement, differences across cultures in self-other agreement have been due to differences in self-ratings and not due to changes in other-ratings (Gentry, Braddy, Fleenor, & Howard, 2008; Gentry, Yip, & Hannum, 2010).

Measurement Characteristics. A few characteristics of the measurement instrument itself have been correlated to increased self-other agreement. Self-raters that have been provided with comparative information about their peers' performance tend to have increased validity of their self-ratings (i.e., their correlation with objective performance criteria; Farh & Dobbins, 1989). Furthermore, when self-ratings and other-ratings are made using a common frame of reference there is more congruence in their ratings (Farh, Werbel, & Bedian, 1988; Steel & Ovalle, 1984).

Wohlers and London (1989) found that there is higher self-other agreement on scales whose managerial behavior items were observable and clearly defined. Felson (1981) found that football players were more likely to over-rate themselves on ambiguous abilities (e.g., mental toughness) than on more concrete ones (e.g., strength). Similarly, Rothermund, Bak, and Brandtstädter (2005) found that students rated themselves more highly on dimensions that were less controllable (e.g., interest in science) than they did on controllable characteristics (e.g., class attendance).

Outcomes of Self-Other Agreement.

Ashford's (1993) theory included self-other agreement as an outcome of the self-appraisal process only, but more recent theories have suggested that there are other potential outcomes associated with this congruence. It is important to note here that comparing self- and other-ratings does not assert that other-ratings are the "true" score and that variability between the two is simply error. While it has been shown that self-ratings tend to be less valid than other-ratings (Harris & Schaubroeck, 1988), it is important in most work settings to have a working knowledge of how others perceive oneself. That is, many important individual outcomes are not determined solely by the individual's performance but indeed by how their performance is perceived by others (e.g., a supervisor's perception of one's performance likely influences promotions, raises, etc.). Atwater and Yammarino (1997) posited many possible outcomes of self-other agreement—some of them have garnered much research interest (e.g., job performance, goal setting), others have gotten very little (e.g., aspiration level), and some unexpected variables have been examined (e.g., derailment).

Job Performance. Much of the early work on self-other agreement focused on the relationship between rating congruence and the performance of those getting feedback. Early results suggest that individual contributors and leaders that have a higher level of agreement with their raters tend to be better performers and are more successful and effective on the job than those who disagree with others' ratings (Atwater & Yammarino, 1992; Bass & Yammarino, 1991; Flocco, 1969; Furnham & Stringfield, 1994; Jennings, 1943; Sosik & Megerian, 1999; Williams & Leavitt, 1947). Some more recent results suggested that there was no relationship between self-other agreement and leader performance (Fleenor, McCauley, & Brutus, 1996; Atwater, Ostroff, Yammarino,

& Fleenor, 1998; Brutus, Fleenor, & Tisak, 1999). However, these studies used dated analytical techniques (e.g., the categorization method) and/or used somewhat flawed logic (e.g., using supervisor ratings as "true" scores).

Investigations without these flaws have yielded positive results. Atwater, Waldman, Ostroff, Robie, and Johnson (2005) used a supervisory rating as the outcome variable in a series of polynomial regressions, but they used a different supervisor than the one that provided the "other" rating as well as a separate scale. Ostroff, Atwater, and Feinberg (2004) used objective criteria (e.g., compensation and organizational level) in addition to a second rating by a supervisor. Results from these studies suggest that there is a positive relationship between self-other agreement and job performance outcomes studied at one point in time.

Some researchers have also examined the effect of self-other agreement on improvement in later performance after receiving feedback, not just a static level of performance. Smither, London, Vasilopoulos, Reilly, Millsap, and Salvemini (1995) examined the effects of upward feedback (i.e., feedback from subordinates to their manager) on subsequent managerial performance. They found that managers who had low to moderate initial performance levels tended to increase in their performance ratings—except those who had agreed with their subordinates on their low performance. Similarly, Atwater, Roush, and Fischthal (1995) found that followers' ratings of their managers' performance increased (after 18 weeks) following the manager's receipt of feedback such that leaders that over-rated themselves improved but those that under-rated themselves did not. Johnson and Ferstl (1999) confirmed these findings once again with a sample of accounting firm managers that were given upward feedback, after a one year

gap between time periods. Walker and Smither (1999) found that, after a five year period, managers improved their multisource feedback ratings most when they met with their direct reports to discuss the previous years' ratings.

Assessment Center Performance. Assessment center performance (AC; Rupp et al., 2015) has been used as the criterion on several occasions to estimate the effect that self-other agreement might have on performance without reliance on ratings. Nowack (1997) found that assessors rated over-estimators and in-agreement/good raters higher on overall AC performance than they did under-estimators and in-agreement/bad raters. Similar results were found for the in-basket task performance as well. Atkins and Wood (2002) conducted a series of polynomial regressions that predicted AC performance from self-ratings and others' (supervisors, peers, and a combined other) ratings of managerial effectiveness. Results suggested that supervisors accurately rated over-estimators (i.e., correctly rated them as average when they rated themselves above average) but underrated under-estimators (i.e., the supervisor and target agreed that their performance was low, but their AC performance was high), while peers tended to over-rate across all levels of AC performance.

Performance in a developmental assessment center (DAC)—an assessment center administered to change an employee's behavior, rather than to predict managerial success (Jones & Whitmore, 1995)—has also been examined as it relates to self-other agreement. Woo, Sims, Rupp, and Gibbons (2008) found that participants who over-estimated their DAC performance relative to their assessor were less behaviorally engaged (i.e., active participation in the DAC, as rated by the assessor) in the developmental program than

were under-raters. There was no correlation, however, between in-agreement pattern (i.e., high and low) and behavioral engagement.

Job Attitudes. Atwater and Yammarino (1997) posited that various job attitudes (e.g., job satisfaction, positive affect) would be influenced by self-rating accuracy. Specifically, they suggested that individuals that over-rate themselves would tend to experience negative attitudes and that those in agreement would tend to experience positive attitudes. Sosik (2001) tested this hypothesis directly by assessing the difference in organizational trust and organizational commitment of supervisors who were categorized according to their self-subordinate rating agreement. They found that supervisors who were in-agreement with their subordinates' ratings of their leadership performance had more trust than both under-estimators and over-estimators. In addition, results suggested that in-agreement raters and (unexpectedly) over-estimators had higher levels of organizational commitment than did under-estimators. While Atwater and Yammarino (1997) theorized only about the job attitudes of the self (i.e., the manager), many researchers have examined the job attitudes of the subordinate.

Szell and Henderson (1997) found that self-subordinate agreement was related to both increased job satisfaction (specifically with supervision, growth, and social aspects of the job) and increased organizational commitment (specifically with identification with the organization) of the subordinate. Sosik (2001) examined the idea that self-subordinate agreement on leadership ratings would influence organizational commitment and trust. He found a significant effect for agreement and subordinate trust: subordinates of over-estimators had the least amount of trust, followed by those of in-agreement raters, while

subordinates of under-estimators had the most trust in their organization. He failed to find a significant effect for organizational commitment.

Similarly, Moshavi, Brown, and Dodd (2003) found that subordinates of under-estimators were more satisfied with their jobs and supervisors than subordinates that were in-agreement with their supervisor; both groups were more satisfied with their jobs and supervisors than those supervised by over-estimators. They also found that subordinates of under-estimators and in-agreement raters reported that they were more productive than those subordinate to over-estimators. However, they did not find an effect of self-other agreement on intent to leave the company.

There is also evidence to suggest that these effects are relevant outside of supervisor-subordinate relationships. Godshalk and Sosik (2000) found that self-other agreement in mentor-mentee relationships (about the mentor's transformational leadership abilities) was positively related to mentee perceptions of the quality of the relationship—mentees of under-estimators experienced the highest levels of psychosocial support, career development, and mentoring effectiveness. Sosik and Godshalk (2004) successfully replicated these results.

Promotion. A few researchers have examined promotions, or rated promotability, as a potential outcome of self-other agreement. This comes directly from Atwater and Yammarino (1997) as well—they hypothesized that both over-estimators and under-estimators would "dramatically misdiagnose strengths and weaknesses and make very poor job-relevant decisions" (p. 158-159). Bass and Yammarino (1991) used a criterion that consisted of a combination of a supervisor's performance rating and their recommendation (or lack thereof) for early promotion of a sample of naval officers. They

found that officers that were in-agreement with their subordinates were more successful than those that were not in agreement. Later, Atwater & Yammarino (1992) used the same criterion measure with naval academy students and found similar results: the correlations between leadership behavior and performance were highest for officers who agreed with their subordinates' ratings of their leadership behavior. Rather than a composite including promotion recommendation ratings or promotability ratings themselves, Halverson, Tonidandel, Barlow, and Dipboye (2002) examined promotion rate directly. They found, through polynomial regression, that self-subordinate agreement on leadership ratings was the most predictive of promotion rate compared to self-peer and self-supervisor agreement: those that were high in agreement also had the highest promotion rate.

Derailment. The Center for Creative Leadership ignited research into managerial derailment—involuntary failure (i.e., plateau in performance, demotion, or firing) of high level executives in the eyes of the organization (McCall & Lombardo, 1983). Initial researchers focused on exploring why derailment happened; a later group investigated the base rate of self-other agreement on derailment behaviors. They found that discrepancies between self-ratings and other-ratings of derailment behaviors occurs in many cultures in studies of European managers (Gentry, Ekelund, Hannum, & Jong, 2007) Hispanic managers (Gentry, Braddy, Fleenor, & Howard, 2008), and Chinese managers (Gentry, Yip, & Hannum, 2010). Tang, Dai, and De Meuse (2010) investigated whether self-other agreement on derailment behavior could predict managerial effectiveness. Managers who under-rated themselves on derailment behavior were found as less effective that those that over-rated themselves, and those that were in-agreement about their high ratings

were less effective than in-agreement low raters. Subsequently, Braddy, Gooty, Fleenor, and Yammarino (2014) examined the relationship between self and other (direct report, peer, and supervisor) ratings of leadership behavior to predict derailment potential. They found that: (1) peer ratings of leader behavior were the most predictive of derailment potential, and (2) derailment is least likely to occur when self-ratings are lower than other-ratings and when self- and other-ratings converge on higher ratings of leader behaviors.

Present Study

The following section will begin with an outline of the major and minor overarching contributions that the results of the present study will make to the self-other agreement and feedback literatures. In addition, the three research questions that are addressed through the course of this dissertation will be introduced, along with the rationale and a short review of the relevant literatures to each question.

Major Contributions

In the previous section I have outlined evidence that suggests that self-other agreement in performance ratings can be influenced by both person and situation variables and that self-other rating congruence is associated with positive individual work outcomes. However, there are certain areas that researchers have not yet addressed for which the present study will provide some major contributions. First, the vast majority of the research in self-other agreement has been conducted on upward feedback (i.e. feedback from subordinate to a supervisor) while downward feedback (i.e. supervisory ratings of subordinate performance) has been largely ignored, except for its use as a criterion. Another major contribution of the present study stems from its longitudinal

design—although some researchers have examined time-lagged effects of feedback (e.g. Smither, London, & Reilly, 2005), only one study was located that used a longitudinal design (i.e. collection of data at three or more time points) to understand the effects of feedback and self-other agreement (Reilly, Smither, & Vasilopoulos, 1996). As a result of this design, modern analytical techniques are used here to answer three basic questions: (1) What is the nature of self-supervisor agreement of performance ratings over time?, (2) What impact does self-supervisor agreement have on important individual level work outcomes?, and (3) Does receipt of other forms of performance feedback impact self-supervisor agreement?

In addition to answering these major questions, the present study makes the following minor contributions to the literature: it provides evidence from a multi-year longitudinal study; it provides data from a mixed sample of both people leaders and individual contributors; it uses modern methodology and data analysis techniques; and it uses objective performance-related criteria (i.e. not performance rating outcomes). Each of these avenues of research has limited empirical evidence available.

Research Questions

The Nature of Agreement. There is some evidence to suggest that individuals tend to increase in agreement with other raters of their performance over time. Reilly et al. (1996) found that the mean difference between self and subordinate ratings of managers' performance from a variety of industries and organizations decreased across four time periods, with one year between each wave. However, they did not examine the relationship between self and supervisory ratings of performance, as they used supervisory ratings as the criterion. Similarly, Bailey and Fletcher (2002) sampled

managers from an automotive services organization that participated in a mandatory multisource feedback procedure (receiving feedback from their supervisor and two subordinates) two years apart. They found that the correlation between self and subordinate ratings increased and the mean difference between self and subordinate ratings decreased—together suggesting an increase in rating congruence—between the two time periods. It should be noted, however, that the sample size was very modest (N = 34). Therefore, there is no direct evidence that speaks to the nature of the relationship of self-ratings with *supervisory* ratings over time. Thus, the present study addresses the following questions regarding the nature of self-supervisor agreement on annual performance ratings:

1) What is the general trend of self-manager agreement on annual performance feedback (i.e., do employees tend to agree or disagree with their managers, and in what direction)?
2) What does this trend look like over time (i.e., do employees tend to diverge from or converge with their managers' ratings)?

Outcomes of Agreement. After the nature of the relationship between self-and supervisor ratings has been elucidated, an understanding of what effects, if any, this agreement has on important individual level outcomes will be investigated. In terms of its effect on job performance, the vast majority of studies to date have used supervisory ratings of performance as the criterion, using self-subordinate or self-peer agreement as the predictor (e.g. Atwater, Ostroff, Yammarino, & Fleenor, 1998). Only two studies were located which used more objective criteria as indicators of job performance. Ostroff, Atwater, and Feinberg (2004) used compensation level (i.e. salary) and organizational

level as outcome variables in polynomial regressions of self-other agreement in three "other" categories: peer, subordinate, and supervisor. They found that self-other agreement was related to these outcome variables in expected ways: (1) agreement on high levels of behavior was related to higher compensation and higher organizational levels than at low levels of agreement, (2) over-estimators had higher compensation and organizational levels than under-estimators, and (3) compensation was lower as congruence between ratings decreased (there was no significant effect for organizational level). Halverson et al. (2002) used a similar methodology to examine the effects of self-other agreement on the promotion rate of US Air Force personnel. They found that only self-subordinate agreement was significantly related to a measure of promotion rate and timeliness. It is possible, however, that the lack of effect for self-supervisor agreement in this case may be due to the fact that the ratings were made of leadership performance. This would, in turn, potentially be most accurately assessed by the followers themselves and therefore most related to promotions. Similarly, both of these studies were conducted at one point in time (or in the case of Halverson et al., 2002, made a longitudinal variable into a static metric). One question that remains unanswered is how self-supervisor agreement may affect the trajectory of these variables over time. Thus, the present study addresses:

3) How does annual performance agreement impact important longitudinal employee-level outcomes (e.g., salary, merit-based pay raises, promotions, organizational level) in terms of:
    a) *Intercept effects* (e.g., do individuals that start at higher organizational levels tend to agree with their manager more?)

b) *Slope effects* (e.g., do individuals that agree with their manager more tend to experience steeper salary trajectories than those that disagree?)

Relatedly, researchers have not yet investigated the effects of self-supervisor agreement on certain dichotomous outcomes like turnover or supervisor change of the individual receiving downward feedback. The closest researchers have come is investigating job attitudes (e.g. satisfaction, commitment) which may in turn be related to outcomes like turnover. However, most researchers have investigated the effects of self-subordinate agreement on the job-related attitudes of the *subordinate* (e.g., Szell & Henderson, 1997). That is, they typically study the job attitudes of the rater rather than the attitudes of the ratee. One exception was located: Sosik (2001) found that increased self-subordinate agreement on ratings of charismatic leadership was associated with increased trust and organizational commitment of the manager (i.e., the self rater). Thus, it appears that agreement may influence certain job attitudes, which may in turn affect behaviors like turnover. Thus, the following question is addressed:

4) How does self-manager agreement relate to important dichotomous outcomes (i.e., turnover, changing supervisors)?

Effects of Feedback. Feedback has been demonstrated to be an effective tool at influencing future job-related behavior, especially at future job performance (e.g. Kluger & DeNisi, 1996). It has also been demonstrated that previous participation in a feedback session can influence rating behavior in future sessions with the same tool. For example, as previously discussed, Bailey and Fletcher (2002) demonstrated that self-other agreement can increase between administrations of the same instrument. However, it remains unclear what effect that receipt of feedback has on rating behavior in other rating

exercises. In other words, does receiving feedback in one exercise make a rater more aware of their performance, in general, and therefore rate themselves more closely to how others rate them in other exercises? Thus:

5) How does taking part in a 360 feedback session affect subsequent agreement on the annual performance feedback?

In addition, there is a substantial amount of evidence to suggest that the effects of self-other agreement vary considerably based on the relative relationship of the self to the other. For example, in their meta-analysis Smither, London, and Reilly (2005) found that the longitudinal effects of feedback on performance were strongest for feedback from direct reports and peers, weak for feedback from supervisors, and basically non-existent for feedback from the self. Therefore, it is reasonable to suspect that the effects of self-other agreement within one feedback tool may affect self-other agreement on another to differing degrees:

a) Does agreement level within the 360 assessment itself moderate this relationship?
b) If so, are there certain groups for which agreement in the 360 assessment is more predictive of future agreement?

Method

Participants

Historical annual performance, annual salary, pay raise, and organizational level data were taken from a sample of 4,630 employees from a large, Midwestern member-owned agricultural cooperative for the years 2011, 2012, 2013, 2014, and 2015. Of these, 3,541 individuals were included in the longitudinal data analysis—the others were

eliminated because they had fewer than three time periods of rating data. Employees consisted of both managers and individual contributors. Multisource feedback data come from a sample of 172 employees.

Measures

Annual Performance Review. The annual review process at this company requires that the individual employee rate themselves on their previous year's performance and their managers rate them for the previous year's performance on a 5-point scale, with both ratings happening around the same time in January. The behavioral anchors for each rating value are included in Table 1. Managers have access to the employees' self-ratings before they conduct their ratings. Performance discussion meetings are scheduled to happen between manager and employee sometime in March.

Competency Ratings. In order to justify the use of the single-item performance rating, individual competency ratings from the year 2015 were used. Employees are rated on a 5-point scale for each competency in a process that is separate from the annual performance review, but around the same time of year. The competency model was developed by Korn Ferry in 2014 on the basis of a series of interviews with senior leadership members of the cooperative. It includes typical performance dimensions such as decision-making, innovation, and productivity. The complete list of performance dimensions and their descriptions can be found in Table 2.

360 Feedback Tool. The multisource feedback tool is primarily used as a developmental procedure within a larger leadership development program of higher performing, high potential employees, but is also occasionally used with low performing employees whose manager wants to provide them developmental feedback. However,

regardless of the context of the survey, its responses are used for developmental purposes only. Its items evaluate similar dimensions to those in the leadership competency model. All items are rated on a six-point Likert scale by a minimum of three other raters from four rater categories: managers, peers, direct reports, or others (e.g. customers or clients). A complete list of items is available in Appendix A. Not all participants received the same subset of items.

Individual Outcome Variables. There are a total of 6 outcome variables in this study: 4 are continuous and 2 are dichotomous in nature. The four continuous variables' (salary, organizational level, pay raise rate, and promotion rate) data were collected for all participants, beginning from 2011. Salary data was simply the individual's numeric salary in USD at any point in time, but for the purposes of constructing trajectories, the natural log of the salary variable was used. Pay raise and promotion data were simply the cumulative number of times that an individual experienced that event; as such, these values were square-rooted to create trajectories. For these transformed variables, the estimated intercept values were re-transformed to their initial metric to aid in interpretation. Organizational level at this organization can vary from 1 to 53, with the majority of the sample working between levels 5 and 15. The process by which individual trajectories and intercepts were modelled is described below. The two dichotomous variables in this study were employee turnover and change in supervisor that was not due to supervisor promotion or supervisor turnover. Each individual was given a value that corresponded to whether or not they had experienced the event in the course of their employment.

Analyses

The Nature of Agreement. In order to describe the nature of longitudinal agreement (i.e. answer the second research question) two linear mixed-effects models were constructed: one that regressed self-ratings of job performance on a time variable and another that regressed managerial ratings of performance on the time variable. Below is a formal specification of the model for the *i*th time point for the *j*th individual:

$$Performance_{ij} = \gamma_{00} + \gamma_{10}(Time_{ij}) + u_{0j} + u_{1j}(Time_{ij}) + e_{ij}$$

As a result, individuals had four data points: the slope and intercept of their performance trajectory as rated by themselves and the slope and intercept of their performance trajectory as rated by their manager. In order to categorize individuals according to their agreement, confidence intervals were constructed around each of these values using the standard error of these estimates. Individuals could rate themselves higher, lower, or equal to their mangers for both their intercepts and their slopes. Thus, an individual could fit into one of nine categories of agreement on the basis of these differences (see Table 3).

Outcomes of Agreement. Individual intercepts and slopes for the four continuous organizational outcomes were constructed in an identical manner to the performance trajectories as described above (i.e. with each outcome regressed onto a time variable in a linear mixed effects model). Variables were transformed as appropriate (e.g. the square root of number of merit-based pay raises was taken, given that it is a count variable) before putting them into the mixed effects model. The individual slope and intercept values were then calculated and converted back into interpretable metrics if the outcome had been transformed.

Then, the outcome slopes and intercepts were used as the outcome variable in a series of regressions, where the outcome slope and intercept were separately regressed first on the linear combination of the self-rating and the manager-rating, then on a combination of the ratings and the squared terms of the ratings (i.e. a polynomial regression). In addition, polynomial regression analyses were also separately run with a number of control variables—employee ethnicity, gender, and tenure—included in both the simple linear and polynomial models. Only when the quadratic model provided significantly better fit than the simple linear model were response surfaces analyzed.

For the dichotomous outcomes (supervisor change and turnover) survival analyses using the Kaplan-Meier (1958) estimator were conducted. For the supervisor change, individuals were only included as having changed a supervisor if the supervisor change did not occur due to supervisor turnover. Similarly, for the turnover analysis individuals were only considered to have turned over when they did so for reasons outside of their control (e.g. retirement, divestiture). A baseline model predicting each event was compared to another model (using the Mantel-Haenszel test) which included the categories of agreement as established above to determine whether agreement had any effect on the effect's occurrence.

Effects of Feedback. In order to assess whether participation in a multisource feedback procedure affects future agreement, a series of linear mixed effects models were conducted on the performance appraisal data from those who participated in the program. The baseline model was that which was described in the Nature of Agreement section above; it was applied to both manager and self ratings of performance separately. Then a model was fit which could test for a change in intercept after feedback and a model which

could test for a change in slope after feedback. A model which included terms for both changes in intercept and slope after feedback simultaneously was fit when at least one of these previous models had significantly better fit than the baseline model.

Results

Use of Single Rating

As previously mentioned, this study relies on the use of a single rating of job performance, measured at one point in time, to attempt to capture a whole year's worth of performance data. To attempt to address this issue, a principal component analysis was constructed. Using a single year's worth of data, ratings made by the supervisor on the subordinate's behavior for each of the competencies in the business's competency model were rotated—93% of the variance in competency scores was accounted for by a single component. Single performance ratings were highly correlated with the component scores ($r = 0.81$). Therefore, we believed the single rating of performance to be "good enough" to proceed with the remainder of the analyses.

The Nature of Agreement

To address the first research question, simple descriptive and agreement statistics of the self and managerial ratings were conducted (see Tables 4 and 5). The means for each year's self and manager ratings were relatively high (always near 3.5) but were relatively variable (ratings *SD*s near 0.7). The average difference between self and manager ratings was also relatively low for all five years (largest difference -0.11), but variable (*SD*s near 0.8). Basic agreement statistics also suggested that there was an acceptable level of disagreement in annual performance ratings to proceed with the

analysis ($\rho \sim 0.44$; $\kappa \sim 0.3$) and that point-in-time agreement was highly variable. See Figure 6 for a histogram of each year's differences in ratings.

The longitudinal nature of agreement was also found to have good variability. The largest category represented in this sample was Reformed Overraters (39%)—individuals that initially over-rated themselves but whose slope indicates that they would converge with their manager's ratings. Refomed Underraters were the next most common category (20%), followed by Serial Overraters (17%), Serial Underraters (11%), Consistent Overraters (5%), Consistent Underraters (3%), and Inflated Agreers (3%). The least represented groups were the Consistent Agreers (1%) and Deflated Agreers (1%).

Outcomes of Agreement

Before considering the results from these analyses, it should be noted that the tables containing the numerical results of the linear and polynomial models fit to these data are only readily interpretable for the simple linear results. The quadratic models' results are difficult to interpret in and of themselves—they serve more to generate the response surfaces from which conclusions are drawn. Thus, it is most appropriate to examine the numeric tables only when better fit is *not* achieved by the quadratic model over the linear model. Alternatively, when the quadratic model does fit the data better than the linear model, it is best to consult the corresponding response surface figure.

A response surface can have three shapes: (1) concave, where it is dome shaped, (2) convex, where it is shaped like a bowl, or (3) saddle, which has both upward and downward curves. Examination of the surface along the line of agreement (where X = Y) and along the line of disagreement (where X = -Y; perpendicular to the line of agreement) also helps to illustrate the effect of agreement on the outcome of interest.

Another approach is to examine the four corners of the X-Y plot—the level of the outcome associated with agreement at the high and low levels of the X-Y scale and the level of the outcome associated with disagreement—where one value is high and the other low, and vice-versa. For the following analyses, there are two different sets of X and Y variables, for which we might expect different plot shapes: (1) initial performance ratings (ratings at time 0 for a given individual) and (2) performance rating trajectories (whether ratings of performance increase or decrease at similar rates).

For initial performance rating agreement, if agreement is universally better, we would expect to find a concave surface with the highest values occurring along the line of agreement, with decreasing values as the surface moves away from this line, and with the lowest values occurring at the perfect disagreement corners of the plot (see Figure 7A). However, we would expect agreement to have differential effects on the outcomes based on the level of performance that is agreed upon. Therefore we would expect more of a saddle-like shape surface. Along the line of agreement the outcome is positively linearly related to the outcome—individuals that agree with their manager about high performance have higher outcomes than those that agree about low levels of performance. We might also expect that the surface appear in a negative U-shape along the line of disagreement—outcomes tend to decrease as disagreement becomes more severe (see Figure 7B). Thus, the set of analyses that use initial performance ratings as the predictors answer the question: does the level of performance influence the effect of agreement on outcomes (both point-in-time and trajectories), or is agreement universally better than disagreement?

For performance rating trajectories the interpretation is a bit more complex. It answers the question: Does agreement in performance rating growth (or decline) affect outcome growth (or decline)? In the case in which giving oneself similar rating slopes to the manager (i.e. longitudinal agreement) is best, we would expect a concave plot similar to the first scenario above. Similarly, we might expect an increasing relationship along the line of agreement—suggesting that agreeing about growing performance is associated with higher levels of the outcome than agreeing about declining performance (see Figure 7C).

However, in the case in which convergence over time is the best, we would expect a saddle-shaped plot, with an upward U-shape along the line of disagreement and a downward or flat curve along the line of agreement (see Figure 7D). This is because we know from previous analyses that individuals that have different slopes in their self-ratings than their managers' ratings are more likely to be converging with their managers' ratings than diverging from them. Of individuals that differ from their manager's rating initially and have different slopes, 74% of them are "reformed" raters rather than "serial" raters. Individuals at the corner of the plot where their manager's slope is positive and their own is negative are more likely to be reformed over-raters than they are to be serial under-raters, and individuals at the corner of the plot where their manager's slope is negative while theirs is positive are more likely to be reformed under-raters than serial over-raters.

Continuous outcomes without control variables. For the first outcome variable of interest (salary) two polynomial regressions provided increased fit over the simple linear model: initial agreement predicted both initial salary and salary trajectories.

Figures 8 and 9 illustrate the response surfaces for these results, respectively. Agreement over time, however, did not predict salary trajectories—the quadratic form of the model did not fit significantly better than the simple linear model. The form of the simple linear model suggests that the slope of the self-ratings predicts salary trajectories the best. A summary of all the models fit to the salary data can be found in Table 6.

For the outcome of starting salary, initial agreement is predictive but varies as a function of the performance level upon which there is agreement. The form of the response surface along the line of perfect agreement is curvilinear: this suggests that initial agreement about both high and low performance is associated with high initial salary, while agreement about moderate performance is associated with the lowest initial salaries. Initial disagreement about performance is positively linearly related to initial salary as a function of performance level such that high self-ratings with low manager ratings is associated with higher compensation than is low self-ratings with high manager ratings.

Perhaps more interestingly, salary trajectories were also predicted by initial agreement. Similar to the above results, agreement is not unilaterally associated with positive outcomes, but is variable on the level of performance upon which there is agreement. Examination of the response surface along the line of perfect agreement suggests that initial agreement about low performance is associated with negative salary trajectories and agreement on high levels of performance is associated with the steepest salary trajectories. The surface along the line of perfect disagreement tells an interesting story: initial disagreement over high performance is much more detrimental to an individual's salary trajectory than is initial disagreement over low performance.

Interestingly, initial low-self ratings, if paired with low manager ratings are associated with the lowest (i.e. negative) salary trajectories but if they are paired with high manager ratings, they are associated with the highest salary trajectories.

For promotion rate, only promotion trajectories were used as outcomes because all employees start with no promotions (i.e. all intercepts are theoretically 0). Initial agreement did not predict promotion rate—again, the quadratic model did not achieve significantly better fit than did the simple linear model. The form of the linear model suggests that the initial ratings by the manager are the most predictive of promotion slopes.

Agreement over time, however, did predict promotion rate (response surface is in Figure 10). The response surface along the line of agreement is flat—suggesting that longitudinal agreement is unilaterally predictive of promotion rate regardless of the magnitude of the slopes. Interestingly, promotion rate along this line was among the lowest. Along the line of perfect disagreement, promotion rate is higher for individuals who have higher disagreement over time—individuals that had very steeply positive trajectories while their manager gave them steeply negative trajectories had the highest promotion rate. A summary of the models fit to the promotion data can be found in Table 7.

Pay raise trajectories were also the only outcome used for the pay raise data. Both initial agreement and agreement over time predicted pay raise trajectories (see Figures 11 and 12). For initial agreement, the effect was once again variable as a function of performance level but in the opposite direction as predicted. Agreement at low levels of performance was associated with steeper pay raise trajectories than agreement at

moderate and high levels of performance. The response surface at the line of disagreement was flat and at the bottom of the trajectories, suggesting that disagreement is unilaterally worse for pay raise trajectories. For agreement over time, perfect agreement had an inverted U-shape relationship with pay raise trajectories—individuals who agreed with their managers on relatively constant slopes had higher pay raise trajectories than individuals agreeing on increasingly positive or negative slopes. As expected with the convergence hypothesis, individuals with different rating slopes than their managers' had the highest levels of pay raise trajectories. Summaries of the models fit to the pay raise data can be found in Table 8.

For organizational level, all three polynomial regressions predicted significantly more variance than the simple linear model (Table 9). For initial organizational level, the line of agreement has an inverted U-shape—individuals agreeing about high levels of performance had the highest organizational level, followed by individuals agreeing about low levels of performance (Figure 13). The line of disagreement has a slightly linear surface: the lowest organizational levels were associated with individuals who initially rated themselves at low levels of performance while their manager gave them a high rating.

Results for the organizational level trajectories were more counter-intuitive. It appears that the initial level of performance influences organizational level trajectories; however, the surface along both the line of agreement and line of disagreement had U-shaped curves, suggesting that perfect agreement at high and low levels of performance are associated with high organizational level, as well as perfect disagreement in both conditions (e.g. whether the self is a 5 and the manager is a 1 or vice-versa). Interestingly,

the line with the minimum value appears to occur where the manager's rating is ~3, and is flat along the self-rating line, suggesting that individuals that receive an initial performance rating near the midpoint of the scale from their manager experience the lowest progression throughout the company (Figure 14).

Organization level trajectories predicted by longitudinal agreement results were similar. Consistent with the convergence hypothesis, the surface along the line of disagreement was U-shaped: individuals with different slopes than their manager's had higher outcomes than individuals with more similar slopes. Consistent with the hypothesis that agreement about increasing trajectories is better than agreement about decreasing trajectories, the surface at the high end of the line of agreement is higher than the surface at the low end of the line—however, the line is U-shaped as well, suggesting that agreeing about decreasing performance is associated with higher organizational level trajectories than agreeing about steadier performance (Figure 15).

Continuous outcomes with control variables. For many of the analyses, including the control variables did very little to alter the results. Initial rating agreement predicted initial salary (Figure 16; Table 10); longitudinal agreement predicted promotion rate trajectories (Figure 17; Table 11); initial performance rating agreement and longitudinal agreement predicted pay raise trajectories (Figures 18 and 19; Table 12); and results for organizational level outcomes (Figures 20, 21, and 22; Table 13) in identical ways, albeit to somewhat less extreme degrees. The only difference in results is for the salary outcomes—without the control variables salary trajectories were predicted by initial performance rating agreement, but with control variables they were not. Conversely, without control variables salary trajectories were not predicted by

longitudinal agreement, but with control variables they were.

The response surface for salary trajectories resembles the response surface for initial salary predicted by initial agreement (Figure 23). Like organizational level trajectories, these results are consistent with the hypothesis that agreement about increasing trajectories is better than agreement about decreasing trajectories, the surface at the high end of the line of agreement is higher than the surface at the low end of the line—however, the line is U-shaped as well, suggesting that agreeing about decreasing performance is associated with higher salary trajectories than agreeing about steadier performance. A summary of the results and their consistencies with the various hypotheses is shown in Table 14.

Dichotomous outcomes. The base rates of each event occurring in the dataset were very different: turnover was very low (3.5%) whereas manager change was much more probable (46%). It is not surprising, then, that turnover was not significantly predicted by longitudinal agreement. Change in manager that is not due to supervisor turnover or promotion, however, was predicted by agreement category (see Table 15 for observed versus expected breakouts for each category). Consistent Under-raters and Reformed Under-raters turned over significantly less often than expected. Serial Overraters turned over at a significantly higher rate than would be expected. To a lesser, non-significant degree, Deflated Agreers turned over more than expected. Similarly, Serial Under-raters turned over less often than expected to a lesser degree.

Effects of Feedback

The results from the analyses that tested whether participation in a feedback program would influence future performance rating trajectories were interesting. For self-

ratings of performance, participation in a multisource feedback procedure did not influence trajectories or slopes—the models including those terms separately did not achieve better fit to the data than the baseline model. However, both models testing the intercept and slope effects on managerial ratings separately and the model which modeled them together fit the data significantly better than the baseline. An examination of the estimates (see Table 16) suggests that following an employee's participation in a multisource feedback process managers' ratings of performance *decrease* in both intercept and slope.

Discussion

Conclusions

There are several conclusions that can be drawn from the results of this research. First, individuals tend to begin their performance rating career either over-rating themselves or under-rating themselves relative to their manager's rating of their performance. This finding is consistent with previous research which suggests that individuals tend to differ from their managers in performance ratings (Fleenor, McCauley, & Brutus, 1996). In addition, these results suggest that the majority of individuals tend to converge with their managers' ratings over time, similar to the findings of Bailey and Fletcher (2002).

In terms of its relationship to individual outcomes, initial agreement is not unilaterally better: as expected, the level of performance in the initial performance ratings upon which there is agreement is a factor in its relationship to outcomes. Consistent with previous research (Brief, Aldag, & Van Sell, 1977; Brutus, Fleenor, & McCauley, 1999; Gentry, Hannum, Ekelund, & de Jong, 2007; Ostroff et al., 2004; Sala, 2003), individuals

that are organizationally higher and have higher salaries are associated with high levels of agreement about initial high performance. In addition, initial agreement can impact important individual longitudinal outcomes: initial agreement predicted salary trajectories, organizational level trajectories, and pay raise trajectories. In the longitudinal outcome case, convergence is important as well for many variables. Agreement over time (i.e., similar slopes) is not associated with high levels of individual outcomes—in fact, disagreement longitudinally is. As previously discussed, when individuals disagree with their manager they are likely *converging* in their ratings (59% of the sample converged in their ratings; only 28% of the disagreeing sample was diverging from each other).

Including control variables (gender, tenure, ethnicity) did little to affect these results. Additionally, agreement was not associated with turnover behavior, but did predict changing managers (not due to manager promotion or turnover) such that those that overrate themselves turnover at a higher rate than expected. Finally, participating in other feedback programs may not positively influence future agreement. It only seems to affect manager ratings, and in a negative way.

Limitations

These analyses, however, are not without their limitations and considerations. First, linear mixed effects models have assumptions that may be considered undesirable. For example, the random effects (i.e., the individual slopes and intercepts) are assumed to be normally distributed. Also, the confidence intervals around the individual slope and intercept values were fairly narrow—so while the intercepts and slopes were statistically different, the magnitude of the differences were sometimes minor. However, all current procedures for estimating individual growth curves (e.g., Latent Curve Analysis) suffer

from these issues. A few more pertinent issues with these particular data are with the polynomial regression interpretation—extrapolation and variance captured. In all response surface analyses, the easiest way to understand their effects are by looking at the four extreme corners of the plot, where X and Y are most extreme and are either equal or opposite. Unfortunately, these parts of the plot are where the least amount of data exist—scatterplots of the available data for both sets of predictors are in Figure 23. Similarly, in many cases, the polynomial models captured significantly more variance than the simple linear model, but in some cases the absolute amount of variance predicted is very low—clearly there is much more at play in the prediction of these outcomes that has not been captured in this study. Thus, most of these results may not replicate in future samples, and the results with particularly low variance explained (e.g., for pay raise data) should be taken with a larger grain of salt.

In addition to the weaknesses of the method and the data, there are certain weaknesses in the design of this research. The most obvious threat to the external validity of these results is that the data come from one single organization, whose business structure is somewhat unique (i.e., a cooperative). However, despite these nominal differences between this business and others, the nature of the work and therefore of perceived performance is not fundamentally different from any other organization, at least of which the author has been a part. In addition, in order to be included in the analyses, each of the individuals had to have been employed at the company for at least 3 years so that they would have at least three data points from which to construct performance rating trajectories. At this company, the majority of individuals for which data were available (76%) met this criterion. However, employees that stay at a company

for more than 2 years are uncommon in the current employment climate—median tenure in 2016 was 4.2 years (U.S. Department of Labor, 2016).

Implications

In conclusion, it seems that performance appraisal agreement can be a useful indicator of underlying organizational behavior. These results provide some evidence that the effects of performance rating agreement can have an impact when it is downward feedback as well, and can impact outcomes longitudinally. Future researchers should capture more explanatory variables (e.g., job codes) and perhaps use both this design but with data from only individuals that started with the system in place (i.e., using only individuals with tenure equal to the study length, that have only ever had one performance appraisal system). There is also a need to replicate these results in other organizational settings. Thus, the goals of this research study were achieved. This is just one study among many possible studies that speak to the nature of agreement over time, the outcomes associated with it, and potential influences of feedback on agreement. In addition, these results provide evidence that performance ratings, with all their flaws and issues, are still a useful tool in the I-O psychologist's toolkit.

References

Adams, S. M. (2005). Positive affect and feedback-giving behavior. *Journal of Managerial Psychology*, *20*(1), 24–42. https://doi.org/10.1108/02683940510571621

Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K., Ollander-Krane, R., & Pulakos, E. D. (2016). Getting Rid of Performance Ratings: Genius or Folly? A Debate. *Industrial and Organizational Psychology*, *9*(2), 219–252. https://doi.org/10.1017/iop.2015.106

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Alder, G. S. (2007). Examining the relationship between feedback and performance in a monitored environment: A clarification and extension of feedback intervention theory. *The Journal of High Technology Management Research*, *17*(2), 157–174. https://doi.org/10.1016/j.hitech.2006.11.004

Alder, G. S., & Ambrose, M. L. (2005). An examination of the effect of computerized performance monitoring feedback on monitoring fairness, performance, and satisfaction. *Organizational Behavior and Human Decision Processes*, *97*(2), 161–177. https://doi.org/10.1016/j.obhdp.2005.03.003

Alimo-Metcalfe, B. (1998). 360 Degree Feedback and Leadership Development. *International Journal of Selection and Assessment*, *6*(1), 35–44. https://doi.org/10.1111/1468-2389.00070

Allen, T. D., & Rush, M. C. (1998). The effects of organizational citizenship behavior on performance judgments: A field study and a laboratory experiment. *Journal of Applied Psychology*, *83*(2), 247. https://doi.org/10.1037/0021-9010.83.2.247

Alvero, A. M., Bucklin, B. R., & Austin, J. (2001). An Objective Review of the Effectiveness and Essential Characteristics of Performance Feedback in Organizational Settings (1985-1998). *Journal of Organizational Behavior Management*, *21*(1), 3–29. https://doi.org/10.1300/J075v21n01_02

Andrews, J. J. W., & Violato, C. (2010). The Assessment of School Psychologists in Practice Through Multisource Feedback. *Canadian Journal of School Psychology*, *25*(4), 328–346. https://doi.org/10.1177/0829573510373585

Anseel, F., Beatty, A. S., Shen, W., Lievens, F., & Sackett, P. R. (2015). How Are We Doing After 30 Years? A Meta-Analytic Review of the Antecedents and Outcomes of Feedback-Seeking Behavior. *Journal of Management*, *41*(1), 318–348. https://doi.org/10.1177/0149206313484521

Anseel, F., & Lievens, F. (2009). The Mediating Role of Feedback Acceptance in the Relationship between Feedback and Attitudinal and Performance Outcomes. *International Journal of Selection & Assessment*, *17*(4), 362–376. https://doi.org/10.1111/j.1468-2389.2009.00479.x

Anseel, F., Lievens, F., & Schollaert, E. (2009). Reflection as a strategy to enhance task performance after feedback. *Organizational Behavior and Human Decision Processes*, *110*(1), 23–35. https://doi.org/10.1016/j.obhdp.2009.05.003

Antonioni, D., & Park, H. (2001). The relationship between rater affect and three sources of 360-degree feedback ratings. *Journal of Management*, *27*(4), 479–495. https://doi.org/10.1016/S0149-2063(01)00104-0

Antonioni, D., & Woehr, D. J. (2001). Improving the Quality of Multisource Rater Performance. In D. Bracken, C. W. Timmreck, & A. H. Church (Eds.), *The handbook of multisource feedback: the comprehensive resource for designing and implementing MSF processes* (1st ed, pp. 114–129). San Francisco: Jossey-Bass.

Araujo, S. V. A., & Taylor, S. N. (2012). The influence of emotional and social competencies on the performance of Peruvian refinery staff. *Cross Cultural Management: An International Journal*, *19*(1), 19–29. https://doi.org/10.1108/13527601211195600

Arthur Jr., W., & Bennett Jr., W. (1995). The International Assignee: The Relative Importance of Factors Perceived to Contribute to Success. *Personnel Psychology*, *48*(1), 99–114.

Ashford, S. J. (1989). Self-Assessments in Organizations: A Literature Review and Integrative Model. *Research in Organizational Behavior*, *11*, 133–174.

Ashford, S. J., & DeStobbeleir, K. E. M. (2013). Feedback, Goal Setting, and Task Performance Revisited. In E. A. Locke & G. P. Latham (Eds.), *New Developments in Goal Setting and Task Performance* (pp. 51–64). New York, NY: Routledge.

Ashford, S. J., & Northcraft, G. B. (1992). Conveying more (or less) than we realize: The role of impression-management in feedback-seeking. *Organizational Behavior*

and Human Decision Processes, *53*(3), 310–334. https://doi.org/10.1016/0749-5978(92)90068-I

Ashford, S. J., & Tsui, A. S. (1991). Self-Regulation for Managerial Effectiveness: The Role of Active Feedback Seeking. *The Academy of Management Journal*, *34*(2), 251–280. https://doi.org/10.2307/256442

Ashton, M. C., & Lee, K. (2007). Empirical, Theoretical, and Practical Advantages of the HEXACO Model of Personality Structure. *Personality and Social Psychology Review*, *11*(2), 150–166. https://doi.org/10.1177/1088868306294907

Atkins, P. W. B., & Wood, R. E. (2002). Self- Versus Others' Ratings as Predictors of Assessment Center Ratings: Validation Evidence for 360-Degree Feedback Programs. *Personnel Psychology*, *55*(4), 871–904. https://doi.org/10.1111/j.1744-6570.2002.tb00133.x

Atwater, L. E., & Brett, J. F. (2005). Antecedents and consequences of reactions to developmental 360° feedback. *Journal of Vocational Behavior*, *66*(3), 532–548. https://doi.org/10.1016/j.jvb.2004.05.003

Atwater, L. E., & Brett, J. F. (2006). 360-Degree Feedback to Leaders: Does it Relate to Changes in Employee Attitudes? *Group & Organization Management*, *31*(5), 578–600. https://doi.org/10.1177/1059601106286887

Atwater, L. E., Brett, J. F., & Charles, A. C. (2007). Multisource feedback: Lessons learned and implications for practice. *Human Resource Management*, *46*(2), 285–307. https://doi.org/10.1002/hrm.20161

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-Other Agreement: Does It Really Matter? *Personnel Psychology*, *51*(3), 577–598.

Atwater, L. E., Waldman, D. A., Atwater, D., & Cartier, P. (2000). An Upward Feedback Field Experiment: Supervisors' Cynicism, Reactions, and Commitment to Subordinates. *Personnel Psychology*, *53*(2), 275–297.

Atwater, L. E., & Yammarino, F. J. (1992). Does Self-Other Agreement on Leadership Perceptions Moderate the Validity of Leadership and Performance Predictions? *Personnel Psychology*, *45*(1), 141–164.

Atwater, L. E., & Yammarino, F. J. (1997). Self-Other Rating Agreement: A Review and Model. *Research in Personnel and Human R*, *15*, 121–174.

Atwater, L., Roush, P., & Fischthal, A. (1995). The Influence of Upward Feedback on Self- and Follower Ratings of Leadership. *Personnel Psychology*, *48*(1), 35–59.

Atwater, L., & Waldman, D. (1998). 360 Degree feedback and leadership development. *The Leadership Quarterly*, *9*(4), 423–426. https://doi.org/10.1016/S1048-9843(98)90009-1

Atwater, L., Waldman, D., Ostroff, C., Robie, C., & Johnson, K. M. (2005). Self–Other Agreement: Comparing its Relationship with Performance in the U.S. and Europe. *International Journal of Selection & Assessment*, *13*(1), 25–40. https://doi.org/10.1111/j.0965-075X.2005.00297.x

Au, A. K. C., & Chan, D. K.-S. (2013). Organizational media choice in performance feedback: a multifaceted approach. *Journal of Applied Social Psychology*, *43*(2), 397–407. https://doi.org/10.1111/j.1559-1816.2013.01009.x

Bailey, C., & Austin, M. (2006). 360 Degree Feedback and Developmental Outcomes: The Role of Feedback Characteristics, Self-Efficacy and Importance of Feedback

Dimensions to Focal Managers' Current Role. *International Journal of Selection and Assessment*, *14*(1), 51–66. https://doi.org/10.1111/j.1468-2389.2006.00333.x

Bailey, C., & Fletcher, C. (2002). The Impact of Multiple Source Feedback on Management Development: Findings from a Longitudinal Study. *Journal of Organizational Behavior*, *23*(7), 853–867.

Baird, L. S. (1977). Self and Superior Ratings of Performance: As Related to Self-Esteem and Satisfaction with Supervision. *The Academy of Management Journal*, *20*(2), 291–300. https://doi.org/10.2307/255402

Balcazar, F., Hopkins, B. L., & Suarez, Y. (1985). A Critical, Objective Review of Performance Feedback. *Journal of Organizational Behavior Management*, *7*(3–4), 65–89. https://doi.org/10.1300/J075v07n03_05

Banker, R. D., Chang, H., & Pizzini, M. J. (2004a). The Balanced Scorecard: Judgmental Effects of Performance Measures Linked to Strategy. *Accounting Review*, *79*(1), 1–23.

Banker, R. D., Chang, H., & Pizzini, M. J. (2004b). The Balanced Scorecard: Judgmental Effects of Performance Measures Linked to Strategy. *Accounting Review*, *79*(1), 1–23.

Barrick, M. R., & Mount, M. K. (1991). The Big Five Personality Dimensions and Job Performance: A Meta-Analysis. *Personnel Psychology*, *44*(1), 1–26.

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology*, *94*(6), 1394. https://doi.org/10.1037/a0016532

Bartels, L. K., & Doverspike, D. (1997). Assessing the assessor: The relationship of assessor personality to leniency in assessment center ratings. *Journal of Social Behavior & Personality*, *12*(5), 179–190.

Bass, A. R., & Turner, J. N. (1973). Ethnic group differences in relationships among criteria of job performance. *Journal of Applied Psychology*, *57*(2), 101. https://doi.org/10.1037/h0037125

Bass, B. M., & Yammarino, F. J. (1991). Congruence of Self and Others' Leadership Ratings of Naval Officers for Understanding Successful Performance. *Applied Psychology*, *40*(4), 437–454. https://doi.org/10.1111/j.1464-0597.1991.tb01002.x

Bell, S. T., & Arthur, W. (2008). Feedback Acceptance in Developmental Assessment Centers: The Role of Feedback Message, Participant Personality, and Affective Response to the Feedback Session. *Journal of Organizational Behavior*, *29*(5), 681–703.

Bernardin, H. J., Cooke, D. K., & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology*, *85*(2), 232. https://doi.org/10.1037/0021-9010.85.2.232

Bernardin, H. J., Thomason, S., Buckley, M. R., & Kane, J. S. (2016). Rater Rating-Level Bias and Accuracy in Performance Appraisals: The Impact OF Rater Personality, Performance Management Competence, and Rater Accountability. *Human Resource Management*, *55*(2), 321–340. https://doi.org/10.1002/hrm.21678

Bernardin, H. J., Tyler, C. L., & Villanova, P. (2009). Rating Level and Accuracy as a Function of Rater Personality. *International Journal of Selection & Assessment*, *17*(3), 300–310. https://doi.org/10.1111/j.1468-2389.2009.00472.x

Bigoness, W. J. (1976). Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. *Journal of Applied Psychology*, *61*(1), 80. https://doi.org/10.1037/0021-9010.61.1.80

Bittner, R. H. (1948). Developing an Industrial Merit Rating Procedure. *Personnel Psychology*, *1*(4), 403–432. https://doi.org/10.1111/j.1744-6570.1948.tb01319.x

Blum, M. L., & Naylor, J. C. (1968). *Industrial Psychology: Its Theoretical and Social Foundations*. New York, NY: Harper & Row.

Bonnel, W., & Boehm, H. (2011). Improving Feedback to Students Online: Teaching Tips From Experienced Faculty. *The Journal of Continuing Education in Nursing*, *42*(11), 503–509. https://doi.org/10.3928/00220124-20110715-02

Bono, J. E., Hooper, A. C., & Yoon, D. J. (2012). Impact of rater personality on transformational and transactional leadership ratings. *The Leadership Quarterly*, *23*(1), 132–145. https://doi.org/10.1016/j.leaqua.2011.11.011

Borman, W. C. (1979). Individual Differences Correlates of Accuracy in Evaluating Others' Performance Effectiveness. *Applied Psychological Measurement*, *3*(1), 103–115. https://doi.org/10.1177/014662167900300111

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology*, *86*(5), 965. https://doi.org/10.1037/0021-9010.86.5.965

Borman, W. C., & Hallam, G. L. (1991). Observation accuracy for assessors of work-sample performance: Consistency across task and individual-differences

correlates. *Journal of Applied Psychology*, *76*(1), 11. https://doi.org/10.1037/0021-9010.76.1.11

Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology*, *80*(1), 168. https://doi.org/10.1037/0021-9010.80.1.168

Bowen, C.-C., Swim, J. K., & Jacobs, R. R. (2000). Evaluating Gender Biases on Actual Job Performance of Real People: A Meta-Analysis1. *Journal of Applied Social Psychology*, *30*(10), 2194–2215. https://doi.org/10.1111/j.1559-1816.2000.tb02432.x

Bracken, D., Timmreck, C. W., & Church, A. H. (Eds.). (2001). *The handbook of multisource feedback: the comprehensive resource for designing and implementing MSF processes* (1st ed). San Francisco: Jossey-Bass.

Braddy, P. W., Gooty, J., Fleenor, J. W., & Yammarino, F. J. (2014). Leader behaviors and career derailment potential: A multi-analytic method examination of rating source and self–other agreement. *The Leadership Quarterly*, *25*(2), 373–390. https://doi.org/10.1016/j.leaqua.2013.10.001

Brett, J. F., & Atwater, L. E. (2001). 360° feedback: Accuracy, reactions, and perceptions of usefulness. *Journal of Applied Psychology*, *86*(5), 930–942. https://doi.org/10.1037//0021-9010.86.5.930

Brockner, J., & Higgins, E. T. (2001). Regulatory Focus Theory: Implications for the Study of Emotions at Work. *Organizational Behavior and Human Decision Processes*, *86*(1), 35–66. https://doi.org/10.1006/obhd.2001.2972

Brutus, S. (2010). Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review*, *20*(2), 144–157. https://doi.org/10.1016/j.hrmr.2009.06.003

Brutus, S., Derayeh, M., Fletcher, C., Bailey, C., Velazquez, P., Shi, K., … Labath, V. (2006). Internationalization of multi-source feedback systems: a six-country exploratory analysis of 360-degree feedback. *International Journal of Human Resource Management*, *17*(11), 1888–1906. https://doi.org/10.1080/09585190601000071

Brutus, S., & Facteau, J. (2003). Short, Simple, and Specific: The Influence of Item Design Characteristics in Multi-Source Assessment Contexts. *International Journal of Selection & Assessment*, *11*(4), 313–325. https://doi.org/10.1111/j.0965-075X.2003.00254.x

Brutus, S., Fleenor, J. W., & McCauley, C. D. (1999). Demographic and personality predictors of congruence in multi-source ratings. *Journal of Management Development*, *18*(5), 417–435. https://doi.org/10.1108/02621719910273569

Brutus, S., Fleenor, J. W., & Tisak, J. (1999). Exploring the Link Between Rating Congruence and Managerial Effectiveness. *Canadian Journal of Administrative Sciences (Canadian Journal of Administrative Sciences)*, *16*(4), 308–322.

Brutus, S., London, M., & Martineau, J. (1999). The impact of 360-degree feedback on planning for career development. *Journal of Management Development*, *18*(8), 676–693. https://doi.org/10.1108/02621719910293774

Caligiuri, P. M. (2000). The Big Five Personality Characteristics as Predictors of Expatriate's Desire to Terminate the Assignment and Supervisor-Rated Performance. *Personnel Psychology*, *53*(1), 67–88.

Campbell, D. J., & Lee, C. (1988). Self-Appraisal in Performance Evaluation: Development Versus Evaluation. *Academy of Management Review*, *13*(2), 302–314. https://doi.org/10.5465/AMR.1988.4306896

Campbell, J. P. (2012). Behavior, Performance, and Effectiveness in the Twenty-First Century. In S. W. J. Kozlowski (Ed.), *The Oxford Handbook of Organizational Psychology, Volume 1*. Oxford University Press. Retrieved from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199928309.001.0001/oxfordhb-9780199928309-e-10

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel Selection in Organizations* (pp. 35–71). San Francisco, CA: Jossey-Bass.

Campion, M. C., Campion, E. D., & Campion, M. A. (2015). Improvements in Performance Management Through the Use of 360 Feedback. *Industrial and Organizational Psychology*, *8*(1), 85–93. https://doi.org/10.1017/iop.2015.3

Cardy, R. L., & Dobbins, G. H. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. *Journal of Applied Psychology*, *71*(4), 672. https://doi.org/10.1037/0021-9010.71.4.672

Carmona, S., Iyer, G., & Reckers, P. M. J. (2014). Performance evaluation bias: A comparative study on the role of financial fixation, similarity-to-self and

likeability. *Advances in Accounting*, *30*(1), 9–17. https://doi.org/10.1016/j.adiac.2014.04.001

Carver, C. S., & Scheier, M. F. (1981). *Attention and Self-Regulation: a Control-Theory Approach to Human Behavior*. New York, NY: Springer New York. Retrieved from http://dx.doi.org/10.1007/978-1-4612-5887-2

Cascio, W. F., & Valenzi, E. R. (1977). Behaviorally anchored rating scales: Effects of education and job experience of raters and ratees. *Journal of Applied Psychology*, *62*(3), 278. https://doi.org/10.1037/0021-9010.62.3.278

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personaltiy Factor Questionnaire (16 PF).* Champaign, IL: Institute for Personality and Ability Testing.

Chen, Y.-Y., & Fang, W. (2008). The Moderating Effect of Impression Management on the Organizational Politics–Performance Relationship. *Journal of Business Ethics*, *79*(3), 263–277. https://doi.org/10.1007/s10551-007-9379-3

Church, A. H. (1997). Do You See What I See? An Exploration of Congruence in Ratings From Multiple Perspectives. *Journal of Applied Social Psychology*, *27*(11), 983–1020. https://doi.org/10.1111/j.1559-1816.1997.tb00283.x

Church, A. H. (2000). Do higher performing managers actually receive better ratings? A validation of multirater assessment methodology. *Consulting Psychology Journal: Practice and Research*, *52*(2), 99–116. https://doi.org/10.1037/1061-4087.52.2.99

Church, A. H., & Bracken, D. W. (1997). Advancing the State of the Art of 360-Degree Feedback Guest Editors' Comments on the Research and Practice of Multirater

Assessment Methods. *Group & Organization Management*, *22*(2), 149–161. https://doi.org/10.1177/1059601197222002

Cleveland, J. N., & Landy, F. J. (1981). The Influence of Rater and Ratee Age on Two Performance Judgments. *Personnel Psychology*, *34*(1), 19–29.

Cleveland, J. N., Lim, A. S., & Murphy, K. M. (2007). Feedback phobia? Why employees do not want to give or receive performance feedback. In J. Langan-Fox, C. L. Cooper, & R. J. Klimoski (Eds.), *Research companion to the dysfunctional workplace: Management challenges and symptoms* (pp. 168–186). Northampton, MA: Edward Elgar Publishing.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology*, *74*(1), 130. https://doi.org/10.1037/0021-9010.74.1.130

Conway, J. M. (1998). Understanding Method Variance in Multitrait-Multirater Performance Appraisal Matrices: Examples Using General Impressions and Interpersonal Affect as Measured Method Factors. *Human Performance*, *11*(1), 29.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric Properties of Multisource Performance Ratings: A meta-Analysis of Subordinate, Supervisor, Peer, and Self-Ratings. *Human Performance*, *10*(4), 331–360. https://doi.org/10.1207/s15327043hup1004_2

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, *13*(6), 653–665. https://doi.org/10.1016/0191-8869(92)90236-I

Coward, W. M., & Sackett, P. R. (1990). Linearity of ability-performance relationships: A reconfirmation. *Journal of Applied Psychology*, *75*(3), 297. https://doi.org/10.1037/0021-9010.75.3.297

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin*, *52*(3), 177–193. https://doi.org/10.1037/h0044919

Curtis, A. B., Harvey, R. D., & Ravden, D. (2005). Sources of Political Distortions in Performance Appraisals. *Group & Organization Management*. https://doi.org/10.1177/1059601104267666

Dai, G., Stiles, P., Hallenbeck, G., & De Meuse, K. R. (2007). *Self-Other Agreement on Leadership Compency Ratings: The Moderating Effects of Rater Perspectives and Rating Ambiguity*. Presented at the Annual Meeting of the Academy of Management, Philadelphia, PA.

David, E. M. (2013). Examining the Role of Narrative Performance Appraisal Comments on Performance. *Human Performance*, *26*(5), 430–450.

Davis, W. D., Carson, C. M., Ammeter, A. P., & Treadway, D. C. (2005). The Interactive Effects of Goal Orientation and Feedback Specificity on Task Performance. *Human Performance*, *18*(4), 409–426.

DeNisi, A. S., & Kluger, A. N. (2000). Feedback Effectiveness: Can 360-Degree Appraisals Be Improved? *The Academy of Management Executive (1993-2005)*, *14*(1), 129–139.

DeNisi, A. S., & Pritchard, R. D. (2006). Performance Appraisal, Performance Management and Improving Individual Performance: A Motivational Framework.

Management and Organization Review, *2*(2), 253–277. https://doi.org/10.1111/j.1740-8784.2006.00042.x

DeShon, R. P., Kozlowski, S. W. J., Schmidt, A. M., Milner, K. R., & Wiechmann, D. (2004). A Multiple-Goal, Multilevel Model of Feedback Effects on the Regulation of Individual and Team Performance. *Journal of Applied Psychology*, *89*(6), 1035–1056. https://doi.org/10.1037/0021-9010.89.6.1035

Dewberry, C., Davies-Muir, A., & Newell, S. (2013). Impact and Causes of Rater Severity/Leniency in Appraisals without Postevaluation Communication Between Raters and Ratees. *International Journal of Selection and Assessment*, *21*(3), 286–293. https://doi.org/10.1111/ijsa.12038

Dobbins, G. H., & Russell, J. M. (1986). The Biasing Effects of Subordinate Likeableness on Leaders' Responses to Poor Performers: A Laboratory and a Field Study. *Personnel Psychology*, *39*(4), 759–777.

Drucker, P. F. (1976). What Results Would You Expect? A Users' Guide to Mbo. *Public Administration Review*, *36*(1), 12.

Duarte, N. T., Goodson, J. R., & Klich, N. R. (1994). Effects of Dyadic Quality and Duration on Performance Appraisal. *Academy of Management Journal*, *37*(3), 499–521. https://doi.org/10.2307/256698

Dunnette, M. D. (1963). A note on the criterion. *Journal of Applied Psychology*, *47*(4), 251. https://doi.org/10.1037/h0040836

Earley, P. C., Northcraft, G. B., Lee, C., & Lituchy, T. R. (1990). Impact of Process and Outcome Feedback on the Relation of Goal Setting to Task Performance. *Academy of Management Journal*, *33*(1), 87–105. https://doi.org/10.2307/256353

Edwards, J. R. (1993). Problems with the Use of Profile Similarity Indices in the Study of Congruence in Organizational Research. *Personnel Psychology*, *46*(3), 641–665.

Edwards, J. R. (1994). The Study of Congruence in Organizational Behavior Research: Critique and a Proposed Alternative. *Organizational Behavior and Human Decision Processes*, *58*(1), 51–100. https://doi.org/10.1006/obhd.1994.1029

Edwards, J. R., & Parry, M. E. (1993). On the Use of Polynomial Regression Equations as an Alternative to Difference Scores in Organizational Research. *Academy of Management Journal*, *36*(6), 1577–1613. https://doi.org/10.2307/256822

Eichinger, R. W., & Lombardo, M. M. (2004). Patterns of Rater Accuracy in 360-Degree Feedback. *Human Resource Planning*, *27*(4), 23–25.

Ellington, J. K., & Wilson, M. A. (2016). The Performance Appraisal Milieu: A Multilevel Analysis of Context Effects in Performance Ratings. *Journal of Business and Psychology*, 1–14. https://doi.org/10.1007/s10869-016-9437-x

Erez, M. (1977). Feedback: A necessary condition for the goal setting-performance relationship. *Journal of Applied Psychology*, *62*(5), 624–627. https://doi.org/10.1037/0021-9010.62.5.624

Facteau, C. L., Facteau, J. D., Schoel, L. C., Russell, J. E. A., & Poteet, M. L. (1998). Reactions of leaders to 360-degree feedback from subordinates and peers. *The Leadership Quarterly*, *9*(4), 427–448. https://doi.org/10.1016/S1048-9843(98)90010-8

Fajfar, P., Campitelli, G., & Labollita, M. (2012). Effects of immediacy of feedback on estimations and performance: Immediacy of feedback and estimations. *Australian*

Journal of Psychology, *64*(3), 169–177. https://doi.org/10.1111/j.1742-9536.2011.00048.x

Fay, C. H., & Latham, G. P. (1982). Effects of Training and Rating Scales on Rating Errors. *Personnel Psychology*, *35*(1), 105–116.

Fedor, D. B. (1991). Recipient responses to performance feedback: A proposed model and its implications. *Research in Personnel and Human Resources*, *9*, 73–120.

Fedor, D. B., Davis, W. D., Maslyn, J. M., & Mathieson, K. (2001). Performance improvement efforts in response to negative feedback: the roles of source power and recipient self-esteem. *Journal of Management*, *27*(1), 79.

Fedor, D. B., & Rensvold, R. B. (1992). An Investigation of Factors Expected to Affect Feedback Seeking: A Longitudinal Field Study. *Personnel Psychology*, *45*(4), 779–805.

Ferris, G. R., Munyon, T. P., Basik, K., & Buckley, M. R. (2008). The performance evaluation context: Social, emotional, cognitive, political, and relationship components. *Human Resource Management Review*, *18*(3), 146–163. https://doi.org/10.1016/j.hrmr.2008.07.006

Ferris, G. R., Russ, G. S., & Fandt, P. M. (1989). Politics in organizations. In R. A. Giacalone & P. Rosenfeld (Eds.), *Impression management in the organization* (pp. 143–170). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Ferris, G. R., Yates, V. L., Gilmore, D. C., & Rowland, K. M. (1985). The Influence of Subordinate Age on Performance Ratings and Causal Attributions. *Personnel Psychology*, *38*(3), 545–557.

Finkelstein, L. M., & Burke, M. J. (1998). Age Stereotyping at Work: The Role of Rater and Contextual Factors on Evaluations of Job Applicants. *The Journal of General Psychology*, *125*(4), 317–345. https://doi.org/10.1080/00221309809595341

Fleenor, J. W., McCauley, C. D., & Brutus, S. (1996). Self-other rating agreement and leader effectiveness. *The Leadership Quarterly*, *7*(4), 487–506. https://doi.org/10.1016/S1048-9843(96)90003-X

Fleenor, J. W., Smither, J. W., Atwater, L. E., Braddy, P. W., & Sturm, R. E. (2010). Self–other rating agreement in leadership: A review. *The Leadership Quarterly*, *21*(6), 1005–1034. https://doi.org/10.1016/j.leaqua.2010.10.006

Fletcher, C., & Baldry, C. (2000). A study of individual differences and self-awareness in the context of multi-source feedback. *Journal of Occupational & Organizational Psychology*, *73*(3), 303–319.

Fletcher, C., Baldry, C., & Cunningham-Snell, N. (1998). The Psychometric Properties of 360 Degree Feedback: An Empirical Study and a Cautionary Tale. *International Journal of Selection and Assessment*, *6*(1), 19–34. https://doi.org/10.1111/1468-2389.00069

Fletcher, C., & Perry, E. L. (2001). Performance Appraisal and Feedback: A Consideration of National Culture and a Review of Contemporary Research and Future Trends. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Visweswaran (Eds.), *Handbook of industrial, work, and organizational psychology* (Vol. 1, pp. 127–144). Thousand Oaks, CA: Sage.

Foster, C. A., & Law, M. R. F. (2006). How Many Perspectives Provide a Compass? Differentiating 360-Degree and Multi-Source Feedback. *International Journal of*

Selection & Assessment, *14*(3), 288–291. https://doi.org/10.1111/j.1468-2389.2006.00347.x

French Jr., J. R. P., & Raven, B. (1959). The Bases of Social Power. In D. P. Cartwright (Ed.), *Studies in Social Power* (pp. 150–167). Ann Arbor, MI: Institute for Social Research, University of Michigan.

Fried, Y., Levi, A. S., Ben-David, H. A., Tiegs, R. B., & Avital, N. (2000). Rater positive and negative mood predispositions as predictors of performance ratings of ratees in simulated and real organizational settings: Evidence from US and Israeli samples. *Journal of Occupational & Organizational Psychology*, *73*(3), 373–378.

Fried, Y., Tiegs, R. B., & Bellamy, A. R. (1992). Personal and interpersonal predictors of supervisors' avoidance of evaluating subordinates. *Journal of Applied Psychology*, *77*(4), 462. https://doi.org/10.1037/0021-9010.77.4.462

Funderburg, S. A., & Levy, P. E. (1997). The Influence of Individual and Contextual Variables on 360-Degree Feedback System Attitudes. *Group & Organization Management*, *22*(2), 210–235. https://doi.org/10.1177/1059601197222005

Furnham, A., & Stringfield, P. (1998). Congruence in Job-Performance Ratings: A Study of 360° Feedback Examining Self, Manager, Peers, and Consultant Ratings. *Human Relations*, *51*(4), 517–530. https://doi.org/10.1177/001872679805100404

Garavan, T. N., Morley, M., & Flynn, M. (1997). 360 degree feedback: its role in employee development. *Journal of Management Development*, *16*(2), 134–147. https://doi.org/10.1108/02621719710164300

Geddes, D., & Konrad, A. M. (2003). Demographic Differences and Reactions to Performance Feedback. *Human Relations, 56*(12), 1485–1513. https://doi.org/10.1177/00187267035612003

Gentry, W. A., Braddy, P. W., Fleenor, J. W., & Howard, P. J. (2008). Self-observer rating discrepancies on the derailment behaviors of Hispanic managers. *The Business Journal of Hispanic Research, 2*(1), 76–87.

Gentry, W. A., Ekelund, B. Z., Hannum, K. M., & Jong, A. de. (2007). A study of the discrepancy between self- and observer-ratings on managerial derailment characteristics of European managers. *European Journal of Work and Organizational Psychology*, *16*(3), 295–325. https://doi.org/10.1080/13594320701394188

Gentry, W. A., Yip, J., & Hannum, K. M. (2010). Self–Observer Rating Discrepancies of Managers in Asia: A study of derailment characteristics and behaviors in Southern and Confucian Asia. *International Journal of Selection & Assessment*, *18*(3), 237–250. https://doi.org/10.1111/j.1468-2389.2010.00507.x

Gillespie, T. L. (2005). Internationalizing 360-Degree Feedback: Are Subordinate Ratings Comparable? *Journal of Business and Psychology*, *19*(3), 361–382. https://doi.org/10.1007/s10869-004-2233-z

Gillespie, T. L., & Parry, R. O. (2006). Fuel for Litigation? Links between Procedural Justice and Multisource Feedback. *Journal of Managerial Issues*, *18*(4), 530–546.

Gioia, D. A., & Longenecker, C. O. (1994). Delving into the Dark Side: The Politics of Executive Appraisal. *Organizational Dynamics*, *22*(3), 47–58.

Godshalk, V. M., & Sosik, J. J. (2000). Does Mentor-Protégé Agreement on Mentor Leadership Behavior Influence the Quality of a Mentoring Relationship? *Group & Organization Management*, *25*(3), 291–317. https://doi.org/10.1177/1059601100253005

Goldsmith, M., & Underhill, B. O. (2001). Multisource Feedback for Executive Development. In D. Bracken, C. W. Timmreck, & A. H. Church (Eds.), *The handbook of multisource feedback: the comprehensive resource for designing and implementing MSF processes* (1st ed, pp. 275–288). San Francisco: Jossey-Bass.

Goldstein, I. L., Emanuel, J. T., & Howell, W. C. (1968). Effect of percentage and specificity of feedback on choice behavior in a probabilistic information-processing task. *Journal of Applied Psychology*, *52*(2), 163. https://doi.org/10.1037/h0025525

Goodman, J. S., & Wood, R. E. (2004). Feedback Specificity, Learning Opportunities, and Learning. *Journal of Applied Psychology*, *89*(5), 809–821. https://doi.org/10.1037/0021-9010.89.5.809

Goodman, J. S., Wood, R. E., & Hendrickx, M. (2004). Feedback Specificity, Exploration, and Learning. *Journal of Applied Psychology*, *89*(2), 248–262. https://doi.org/10.1037/0021-9010.89.2.248

Govaerts, M. J. B., Wiel, M. W. J. van de, & Vleuten, C. P. M. van der. (2013). Quality of feedback following performance assessments: does assessor expertise matter? *European Journal of Training and Development*, *37*(1), 105–125. https://doi.org/10.1108/03090591311293310

Graen, G. B., & Uhl-Bien, M. (1995). Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective. *The Leadership Quarterly*, *6*(2), 219–247. https://doi.org/10.1016/1048-9843(95)90036-5

Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of Race on Organizational Experience, Job Performance Evaluations, and Career Outcomes. *Academy of Management Journal*, *33*(1), 64–86. https://doi.org/10.2307/256352

Greguras, G. J. (2005). Managerial Experience and the Measurement Equivalence of Performance Ratings. *Journal of Business and Psychology*, *19*(3), 383–397. https://doi.org/10.1007/s10869-004-2234-y

Greguras, G. J., Ford, J. M., & Brutus, S. (2003). Manager attention to multisource feedback. *Journal of Management Development*, *22*(4), 345–361. https://doi.org/10.1108/02621710310467631

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology*, *83*(6), 960–968. https://doi.org/10.1037/0021-9010.83.6.960

Hall, F. S., & Hall, D. T. (1976). Effects of Job Incumbents' Race and Sex of Evaluations of Managerial Performance. *Academy of Management Journal*, *19*(3), 476–481. https://doi.org/10.2307/255613

Halverson, S. K., Tonidandel, S., Barlow, C., & Dipboye, R. L. (2002, April). *Self-Other Agreement on a 360-Degree Leadership Evaluation*. Poster presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Canada.

Hammer, M. R., Gudykunst, W. B., & Wiseman, R. L. (1978). Dimensions of intercultural effectiveness: An exploratory study. *International Journal of Intercultural Relations*, *2*(4), 382–393. https://doi.org/10.1016/0147-1767(78)90036-6

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. (1974). Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. *Journal of Applied Psychology*, *59*(6), 705. https://doi.org/10.1037/h0037503

Harackiewicz, J. M. (1979). The effects of reward contingency and performance feedback on intrinsic motivation. *Journal of Personality and Social Psychology*, *37*(8), 1352–1363. https://doi.org/10.1037/0022-3514.37.8.1352

Harari, M. B., Rudolph, C. W., & Laginess, A. J. (2015). Does rater personality matter? A meta-analysis of rater Big Five–performance rating relationships. *Journal of Occupational and Organizational Psychology*, *88*(2), 387–414. https://doi.org/10.1111/joop.12086

Harris, M. M. (1994). Rater Motivation in the Performance Appraisal Context: A Theoretical Framework. *Journal of Management*, *20*(4), 735–756. https://doi.org/10.1177/014920639402000403

Harris, M. M., & Schaubroeck, J. (1988). A Meta-Analysis of Self--Supervisor, Self--Peer, and Peer--Supervisor Ratings. *Personnel Psychology*, *41*(1), 43–62.

Hauenstein, N. M. A., & Alexander, R. A. (1991). Rating ability in performance judgments: The joint influence of implicit theories and intelligence. *Organizational Behavior and Human Decision Processes*, *50*(2), 300–323. https://doi.org/10.1016/0749-5978(91)90024-N

Hazucha, J. F., Hezlett, S. A., & Schneider, R. J. (1993). The impact of 360-degree feedback on management skills development. *Human Resource Management*, *32*(2–3), 325–351. https://doi.org/10.1002/hrm.3930320210

Heidemeier, H., & Moser, K. (2009). Self–other agreement in job performance ratings: A meta-analytic test of a process model. *Journal of Applied Psychology*, *94*(2), 353–370. https://doi.org/10.1037/0021-9010.94.2.353

Heslin, P. A., & Latham, G. P. (2004). The Effect of Upward Feedback on Managerial Behavior. *Applied Psychology*, *53*(1), 23–37. https://doi.org/10.1111/j.1464-0597.2004.00159.x

Heslin, P. A., Latham, G. P., & Walle, D. V. (2005). The Effect of Implicit Person Theory on Performance Appraisals. *Journal of Applied Psychology*, *90*(5), 842–856. https://doi.org/10.1037/0021-9010.90.5.827

Heslin, P. A., & VandeWalle, D. (2008). Managers' Implicit Assumptions About Personnel. *Current Directions in Psychological Science*, *17*(3), 219–223. https://doi.org/10.1111/j.1467-8721.2008.00578.x

Hezlett, S. A. (2008). Using Multisource Feedback to Develop Leaders: Applying Theory and Research to Improve Practice. *Advances in Developing Human Resources*, *10*(5), 703–720. https://doi.org/10.1177/1523422308322271

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280–1300. https://doi.org/10.1037/0003-066X.52.12.1280

Higgins, E. T., Shah, J., & Friedman, R. (1997). Emotional responses to goal attainment: Strength of regulatory focus as moderator. *Journal of Personality and Social Psychology*, *72*(3), 515–525. https://doi.org/10.1037/0022-3514.72.3.515

Hogan, E. A. (1987). Effects of Prior Expectations on Performance Ratings: A Longitudinal Study. *Academy of Management Journal*, *30*(2), 354–368. https://doi.org/10.2307/256279

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology*, *63*(5), 579–588. https://doi.org/10.1037/0021-9010.63.5.579

Hooijberg, R., & Lane, N. (2009). Using Multisource Feedback Coaching Effectively in Executive Education. *Academy of Management Learning & Education*, *8*(4), 483–493.

Huber, V. L., Neale, M. A., & Nofthcraft, G. B. (1987). Judgment by heuristics: Effects of ratee and rater characteristics and performance standards on performance-related judgments. *Organizational Behavior and Human Decision Processes*, *40*(2), 149–169. https://doi.org/10.1016/0749-5978(87)90010-0

Hunter, J. E., & Schmidt, F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, and Law*, *2*(3–4), 447. https://doi.org/10.1037/1076-8971.2.3-4.447

Idson, L. C., & Higgins, E. T. (2000). How current feedback and chronic effectiveness influence motivation: everything to gain versus everything to lose. *European Journal of Social Psychology*, *30*(4), 583–592. https://doi.org/10.1002/1099-0992(200007/08)30:4<583::AID-EJSP9>3.0.CO;2-S

Idson, L. C., Liberman, N., & Higgins, E. T. (2000). Distinguishing Gains from Nonlosses and Losses from Nongains: A Regulatory Focus Perspective on

Hedonic Intensity. *Journal of Experimental Social Psychology*, *36*(3), 252–274. https://doi.org/10.1006/jesp.1999.1402

Ilgen, D., & Davis, C. (2000). Bearing Bad News: Reactions to Negative Performance Feedback. *Applied Psychology*, *49*(3), 550–565. https://doi.org/10.1111/1464-0597.00031

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance Appraisal Process Research in the 1980s: What Has It Contributed to Appraisals in Use? *Organizational Behavior and Human Decision Processes*, *54*(3), 321–368. https://doi.org/10.1006/obhd.1993.1015

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, *64*(4), 349–371. https://doi.org/10.1037/0021-9010.64.4.349

Jawahar, I. M. (2010). The Mediating Role of Appraisal Feedback Reactions on the Relationship Between Rater Feedback-Related Behaviors and Ratee Performance. *Group & Organization Management*, *35*(4), 494–526. https://doi.org/10.1177/1059601110378294

Jawahar, I. M., & Williams, C. R. (1997). Where All the Children Are Above Average: The Performance Appraisal Purpose Effect. *Personnel Psychology*, *50*(4), 905–926.

Johnson, D. S., Perlow, R., & Pieper, K. F. (1993). Differences in Task Performance as a Function of Type of Feedback: Learning-Oriented Versus Performance-Oriented Feedback1. *Journal of Applied Social Psychology*, *23*(4), 303–320. https://doi.org/10.1111/j.1559-1816.1993.tb01089.x

Johnson, J. W., & Ferstl, K. L. (1999). The Effects of Interrater and Self-Other Agreement on Performance Improvement Following Upward Feedback. *Personnel Psychology*, *52*(2), 271–303.

Jones, L., & Fletcher, C. (2002). Self-assessment in a selection situation: An evaluation of different measurement approaches. *Journal of Occupational & Organizational Psychology*, *75*(2), 145–161.

Jones, R. G., & Whitmore, M. D. (1995). Evaluating Developmental Assessment Centers as Interventions. *Personnel Psychology*, *48*(2), 377–388.

Judge, T. A., & Ferris, G. R. (1993). Social Context of Performance Evaluation Decisions. *Academy of Management Journal*, *36*(1), 80–105. https://doi.org/10.2307/256513

Judge, T. A., LePine, J. A., & Rich, B. L. (2006). Loving Yourself Abundantly: Relationship of the Narcissistic Personality to Self- and Other Perceptions of Workplace Deviance, Leadership, and Task and Contextual Performance. *Journal of Applied Psychology*, *91*(4), 762–776.

Jurgensen, C. E. (1950). Intercorrelations in merit rating traits. *Journal of Applied Psychology*, *34*(4), 240. https://doi.org/10.1037/h0062163

Kaiser, R. B., & Craig, S. B. (2005). Building a Better Mouse Trap: Item Characteristics Associated With Rating Discrepancies in 360-Degree Feedback. *Consulting Psychology Journal: Practice & Research*, *57*(4), 235–245. https://doi.org/10.1037/1065-9293.57.4.235

Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J. (1995). Stability of Rater Leniency: Three Studies. *Academy of Management Journal*, *38*(4), 1036–1051. https://doi.org/10.2307/256619

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, *53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Kaplan, R. S., & Norton, D. P. (1992). The Balanced Scorecard--Measures That Drive Performance. *Harvard Business Review*, *70*(1), 71–79.

Kaplan, R. S., & Norton, D. P. (1996). Using the Balanced Scorecard as a Strategic Management System. *Harvard Business Review*, *74*(1), 75–85.

Kaplan, S. E., Petersen, M. J., & Samuels, J. A. (2007). Effects of Subordinate Likeability and Balanced Scorecard Format on Performance-Related Judgments. *Advances in Accounting*, *23*, 85–111. https://doi.org/10.1016/S0882-6110(07)23004-4

Kernan, M. C., Heimann, B., & Hanges, P. J. (1991). Effects of Goal Choice, Strategy Choice, and Feedback Source on Goal Acceptance, Performance and Subsequent Goals1. *Journal of Applied Social Psychology*, *21*(9), 713–733. https://doi.org/10.1111/j.1559-1816.1991.tb00544.x

Kinicki, A. J., Prussia, G. E., Wu, B. (Joshua), & McKee-Ryan, F. M. (2004). A Covariance Structure Analysis of Employees' Response to Performance Feedback. *Journal of Applied Psychology*, *89*(6), 1057–1069. https://doi.org/10.1037/0021-9010.89.6.1057

Klimoski, R., & Inks, L. (1990). Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes*, *45*(2), 194–208. https://doi.org/10.1016/0749-5978(90)90011-W

Kline, T. J. B., & Sulsky, L. M. (2009). Measurement and assessment issues in performance appraisal. *Canadian Psychology/Psychologie Canadienne*, *50*(3), 161–171. https://doi.org/10.1037/a0015668

Klores, M. S. (1966). Rater Bias in Forced-Distribution Performance Ratings. *Personnel Psychology*, *19*(4), 411–421.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kluger, A. N., Lewinsohn, S., & Aiello, J. R. (1994). The Influence of Feedback on Mood: Linear Effects on Pleasantness and Curvilinear Effects on Arousal. *Organizational Behavior and Human Decision Processes*, *60*(2), 276–299. https://doi.org/10.1006/obhd.1994.1084

Kolz, A. R., McFarland, L. A., & Silverman, S. B. (1998). Ognitive ability and job experience as predictors of work performance. *Journal of Psychology*, *132*(5), 539.

Korsgaard, M. A. (1996). The Impact of Self-Appraisals on Reactions to Feedback from Others: The Role of Self-Enhancement and Self-Consistency Concerns. *Journal of Organizational Behavior*, *17*(4), 301–311.

Korsgaard, M. A., & Diddams, M. (1996). The Effect of Process Feedback and Task Complexity on Personal Goals, Information Searching, and Performance Improvement. *Journal of Applied Social Psychology*, *26*(21), 1889–1911. https://doi.org/10.1111/j.1559-1816.1996.tb00104.x

Kozlowski, S. W., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology*, *77*(2), 161. https://doi.org/10.1037/0021-9010.77.2.161

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*, *70*(1), 56. https://doi.org/10.1037/0021-9010.70.1.56

Kulik, J. A., & Kulik, C.-L. C. (1988). Timing of Feedback and Verbal Learning. *Review of Educational Research*, *58*(1), 79–97. https://doi.org/10.2307/1170349

Kuvaas, B., Buch, R., & Dysvik, A. (2016). Constructive Supervisor Feedback is not Sufficient: Immediacy and Frequency is Essential. *Human Resource Management*, n/a-n/a. https://doi.org/10.1002/hrm.21785

Lahuis, D. M., & Avis, J. M. (2007). Using Multilevel Random Coefficient Modeling to Investigate Rater Effects in Performance Ratings. *Organizational Research Methods*, *10*(1), 97–107. https://doi.org/10.1177/1094428106289394

Lam, C. F., DeRue, D. S., Karam, E. P., & Hollenbeck, J. R. (2011). The impact of feedback frequency on learning and task performance: Challenging the "more is better" assumption. *Organizational Behavior and Human Decision Processes*, *116*(2), 217–228. https://doi.org/10.1016/j.obhdp.2011.05.002

Lance, C. E., Hoffman, B. J., Gentry, W. A., & Baranik, L. E. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review*, *18*(4), 223–232. https://doi.org/10.1016/j.hrmr.2008.03.002

Landau, J. (1995). The Relationship of Race and Gender to Managers' Ratings of Promotion Potential. *Journal of Organizational Behavior*, *16*(4), 391–400.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*(1), 72. https://doi.org/10.1037/0033-2909.87.1.72

Larson Jr., J. R. (1986). Supervisors' performance feedback to subordinates: The impact of subordinate performance valence and outcome dependence. *Organizational Behavior and Human Decision Processes*, *37*(3), 391–408. https://doi.org/10.1016/0749-5978(86)90037-3

Latham, G. P., & Dello Russo, S. (2008). The Influence of Organizational Politics on Performance Appraisal. In S. Cartwright & C. Cooper (Eds.), *The Oxford Handbook of Personnel Psychology* (pp. 388–410). Oxford, UK: Oxford University Press.

Latham, G. P., & Mann, S. (2006). Advances in the Science of Performance Appraisal: Implications for Practice. In G. P. Hodgkinson & J. K. Ford (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 21, pp. 295–337).

Latham, G. P., & Wexley, K. N. (1977). Behavioral Observation Scales for Performance Appraisal Purposes. *Personnel Psychology*, *30*(2), 255–268.

Lawler, E. E. (1967). The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology*, *51*(5p1), 369. https://doi.org/10.1037/h0025095

Lefkowitz, J. (2000). The role of interpersonal affective regard in supervisory performance ratings: A literature review and proposed causal model. *Journal of Occupational & Organizational Psychology*, *73*(1), 67–85.

Lepsinger, R., & Lucia, A. D. (2009). *The Art and Science of 360-Degree Feedback* (2nd ed.). San Francisco, CA: Jossey-Bass.

Levy, P. E., & Williams, J. R. (2004). The Social Context of Performance Appraisal: A Review and Framework for the Future. *Journal of Management*, *30*(6), 881–905. https://doi.org/10.1016/j.jm.2004.06.005

Liden, R. C., Stilwell, D., & Ferris, G. R. (1996). The Effects of Supervisor and Subordinate Age on Objective Performance and Subjective Performance Ratings. *Human Relations*. https://doi.org/10.1177/001872679604900304

Locke, E. A. (1991). Goal theory vs. control theory: Contrasting approaches to understanding work motivation. *Motivation and Emotion*, *15*(1), 9–28. https://doi.org/10.1007/BF00991473

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717. https://doi.org/10.1037//0003-066X.57.9.705

Lockyer, J., Armson, H., Chesluk, B., Dornan, T., Holmboe, E., Loney, E., … Sargeant, J. (2011). Feedback data sources that inform physician self-assessment. *Medical Teacher*, *33*(2), e113–e120. https://doi.org/10.3109/0142159X.2011.542519

London, M., & Smither, J. W. (1995). Can Multi-Source Feedback Change Perceptions of Goal Accomplishment, Self-Evaluations, and Performance-Related Outcomes? Theory-Based Applications and Directions for Research. *Personnel Psychology*, *48*(4), 803–839.

London, M., & Smither, J. W. (2002). Feedback orientation, feedback culture, and the longitudinal performance management process. *Human Resource Management Review*, *12*, 81–100.

Longenecker, C. O., Sims, J., Henry P., & Gioia, D. A. (1987). Behind the Mask: The Politics of Employee Appraisal. *Academy of Management Executive (08963789)*, *1*(3), 183–193. https://doi.org/10.5465/AME.1987.4275731

Luthans, F., & Peterson, S. J. (2003). 360-degree feedback with systematic coaching: Empirical analysis suggests a winning combination. *Human Resource Management*, *42*(3), 243–256. https://doi.org/10.1002/hrm.10083

Luthans, K. W., & Farner, S. (2002). Expatriate development: the use of 360-degree feedback. *Journal of Management Development*, *21*(10), 780–793. https://doi.org/10.1108/02621710210448048

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology*, *67*(3), 280–296. https://doi.org/10.1037/0021-9010.67.3.280

Mai-Dalton, R. R., Feldman-Summers, S., & Mitchell, T. R. (1979). Effect of employee gender and behavioral style on the evaluations of male and female banking executives. *Journal of Applied Psychology*, *64*(2), 221. https://doi.org/10.1037/0021-9010.64.2.221

Maurer, T. J., Mitchell, D. R. D., & Barbeite, F. G. (2002). Predictors of attitudes toward a 360-degree feedback system and involvement in post-feedback management development activity. *Journal of Occupational and Organizational Psychology*, *75*(1), 87–107. https://doi.org/10.1348/096317902167667

Maurer, T. J., Palmer, J. K., & Lisnov, S. S. (1995). Distinguishing Context Effects From Context Errors in Judgments of Behavior. *Journal of Applied Social Psychology*, *25*(18), 1637–1651. https://doi.org/10.1111/j.1559-1816.1995.tb02637.x

Mayo, M., Kakarika, M., Pastor, J. C., & Brutus, S. (2012). Aligning or Inflating Your Leadership Self-Image? A Longitudinal Study of Responses to Peer Feedback in MBA Teams. *Academy of Management Learning & Education*, *11*(4), 631–652. https://doi.org/10.5465/amle.2010.0069

McCall, M. W. J., & Lombardo, M. M. (1983). *Off the track: Why and how successful executives get derailed.* (Technical Report No. 21). Greensboro, NC: Center for Creative Leadership.

McCarthy, A. M., & Garavan, T. N. (1999). Developing self-awareness in the managerial career development process: the value of 360-degree feedback and the MBTI. *Journal of European Industrial Training*, *23*(9), 437–445. https://doi.org/10.1108/03090599910302613

McKay, P. F., & McDaniel, M. A. (2006). A Reexamination of Black--White Mean Differences in Work Performance: More Data, More Moderators. *Journal of Applied Psychology*, *91*(3), 538–554.

Mendenhall, M., & Oddou, G. (1985). The Dimensions of Expatriate Acculturation: A Review. *Academy of Management Review*, *10*(1), 39–47. https://doi.org/10.5465/AMR.1985.4277340

Mero, N. P., Guidice, R. M., & Brownlee, A. L. (2007). Accountability in a Performance Appraisal Context: The Effect of Audience and Form of Accounting on Rater Response and Behavior. *Journal of Management*, *33*(2), 223–252. https://doi.org/10.1177/0149206306297633

Mero, N. P., & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology*, *80*(4), 517. https://doi.org/10.1037/0021-9010.80.4.517

Meyer, H. H., Kay, E., & French Jr., J. R. P. (1965). Split Roles in Performance Appraisal. *Harvard Business Review*, *43*(1), 123–129.

Miller, B. K., Rutherford, M. A., & Kolodinsky, R. W. (2008). Perceptions of Organizational Politics: A Meta-analysis of Outcomes. *Journal of Business and Psychology*, *22*(3), 209–222. https://doi.org/10.1007/s10869-008-9061-5

Mobley, W. H. (1982). Supervisor and Employee Race and Sex Effects on Performance Appraisals: A Field Study of Adverse Impact and Generalizability. *Academy of Management Journal*, *25*(3), 598–606. https://doi.org/10.2307/256083

Morgeson, F. P., Delaney-Klinger, K., & Hemingway, M. A. (2005). The Importance of Job Autonomy, Cognitive Ability, and Job-Related Skill for Predicting Role Breadth and Job Performance. *Journal of Applied Psychology*, *90*(2), 399–406. https://doi.org/10.1037/0021-9010.90.2.399

Morgeson, F. P., Mumford, T. V., & Campion, M. A. (2005). Coming Full Circle: Using Research and Practice to Address 27 Questions About 360-Degree Feedback Programs. *Consulting Psychology Journal: Practice and Research*, *57*(3), 196–209. https://doi.org/10.1037/1065-9293.57.3.196

Morrison, E. W., & Weldon, E. (1990). The Impact of an Assigned Performance Goal on Feedback Seeking Behavior. *Human Performance*, *3*(1), 37–50.

Moshavi, D., Brown, F. W., & Dodd, N. G. (2003). Leader self-awareness and its relationship to subordinate attitudes and performance. *Leadership & Organization Development Journal*, *24*(7), 407–418. https://doi.org/10.1108/01437730310498622

Moss, S. E., Valenzi, E. R., & Taggart, W. (2003). Are You Hiding from Your Boss? The Development of a Taxonomy and Instrument to Assess the Feedback Management Behaviors of Good and Bad Performers. *Journal of Management*, *29*(4), 487–510. https://doi.org/10.1016/S0149-2063_03_00022-9

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, Rater and Level Effects in 360-Degree Performance Ratings. *Personnel Psychology*, *51*(3), 557–576. https://doi.org/10.1111/j.1744-6570.1998.tb00251.x

Mount, M. K., & Thompson, D. E. (1987). Cognitive categorization and quality of performance ratings. *Journal of Applied Psychology*, *72*(2), 240. https://doi.org/10.1037/0021-9010.72.2.240

Murphy, K. R. (2008). Explaining the Weak Relationship Between Job Performance and Ratings of Job Performance. *Industrial and Organizational Psychology*, *1*(2), 148–160. https://doi.org/10.1111/j.1754-9434.2008.00030.x

Murphy, K. R., Balzer, W. K., Lockhart, M. C., & Eisenman, E. J. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology*, *70*(1), 72. https://doi.org/10.1037/0021-9010.70.1.72

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? *APA PsycNET*, *67*(5), 562. https://doi.org/10.1037/0021-9010.67.5.562

Naylor, J. C., Pritchard, R. D., & Ilgen, D. R. (1980). *A theory of behavior in organizations*. London ; New York: Academic Press.

Neubert, M. J. (1998). The Value of Feedback and Goal Setting Over Goal Setting Alone and Potential Moderators of this Effect: a Meta-Analysis. *Human Performance*, *11*(4), 321–335. https://doi.org/10.1207/s15327043hup1104_2

Ng, K.-Y., Koh, C., Ang, S., Kennedy, J. C., & Chan, K.-Y. (2011). Rating leniency and halo in multisource feedback ratings: Testing cultural assumptions of power distance and individualism-collectivism. *Journal of Applied Psychology*, *96*(5), 1033–1044. https://doi.org/10.1037/a0023368

Nieva, V. F., & Gutek, B. A. (1980). Sex Effects on Evaluation. *Academy of Management Review*, *5*(2), 267–276. https://doi.org/10.5465/AMR.1980.4288749

Nilsen, D., & Campbell, D. P. (1993). Self–observer rating discrepancies: Once an overrater, always an overrater? *Human Resource Management*, *32*(2–3), 265–281. https://doi.org/10.1002/hrm.3930320206

Northcraft, G. B., Huber, V., & Neale, M. A. (1988). Sex Effects in Performance--Related Judgments. *Human Performance*, *1*(3), 161.

Northcraft, G. B., Schmidt, A. M., & Ashford, S. J. (2011). Feedback and the rationing of time and effort among competing tasks. *Journal of Applied Psychology*, *96*(5), 1076. https://doi.org/10.1037/a0023221

Nowack, K. M. (1997). Congruence between self-other ratings and assessment center performance. *Journal of Social Behavior & Personality*, *12*(5), 145–166.

Nowack, K. M. (2009). Leveraging multirater feedback to facilitate successful behavioral change. *Consulting Psychology Journal: Practice and Research*, *61*(4), 280–297. https://doi.org/10.1037/a0017381

Nowack, K. M., & Mashihi, S. (2012). Evidence-based answers to 15 questions about leveraging 360-degree feedback. *Consulting Psychology Journal: Practice and Research*, *64*(3), 157–182. https://doi.org/10.1037/a0030011

Ogunfowora, B., Bourdage, J., & Lee, K. (2010). Rater Personality and Performance Dimension Weighting in Making Overall Performance Judgments. *Journal of Business and Psychology*, *25*(3), 465–476. https://doi.org/10.1007/s10869-009-9144-y

Oliphant, V. N., & Alexander III, E. R. (1982). Reactions to Resumes as a Function of Resume Determinateness, Applicant Characteristics, and Sex of Raters. *Personnel Psychology*, *35*(4), 829–842.

O'Neil, D. P., Sweeney, P. J., Ness, J., & Kolditz, T. A. (2007). Leader Development and Self-Awareness in the U.S. Army Bench Project. In D. Crandall (Ed.), *Leadership Lessons from West Point* (1st ed., pp. 107–130). San Francisco, CA: Jossey-Bass.

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth-Fidelity Dilemma in Personality Measurement for Personnel Selection. *Journal of Organizational Behavior*, *17*(6), 609–626.

Organ, D. W. (1988). *Organizational citizenship behavior: the good soldier syndrome*. Lexington, Mass: Lexington Books.

Ostroff, C., Atwater, L. E., & Feinberg, B. J. (2004). Understanding Self-Other Agreement: A Look at Rater and Ratee Characteristics, Context, and Outcomes. *Personnel Psychology*, *57*(2), 333–375.

Overeem, K., Lombarts, M. J. M. H., Arah, O. A., Klazinga, N. S., Grol, R. P. T. M., & Wollersheim, H. C. (2010). Three methods of multi-source feedback compared: A plea for narrative comments and coworkers' perspectives. *Medical Teacher*, *32*(2), 141–147. https://doi.org/10.3109/01421590903144128

Paterson, D. G. (1922). The Scott Company Graphic Rating Scale. *Journal of Personnel Research*, *1*, 361–376.

Patiar, A., & Mia, L. (2008). The effect of subordinates' gender on the difference between self-ratings, and superiors' ratings, of subordinates' performance in hotels. *The International Journal of Hospitality Management*, *27*, 53–64.

Pearce, J. L., & Porter, L. W. (1986). Employee responses to formal performance appraisal feedback. *Journal of Applied Psychology*, *71*(2), 211–218. https://doi.org/10.1037/0021-9010.71.2.211

Peiperl, M. A. (2001). Getting 360° Feedback Right. *Harvard Business Review*, *79*(1), 142–147.

Pfau, B., & Kay, I. (2002). does 360-degree feedback negatively affect company performance? *HR Magazine*, *47*(6), 54.

Poon, J. (2001). Mood: A Review of Its Antecedents and Consequences. *International Journal of Organization Theory & Behavior (Marcel Dekker)*, *4*(3/4), 357.

Pritchard, R. D., Harrell, M. M., DiazGranados, D., & Guzman, M. J. (2008). The productivity measurement and enhancement system: A meta-analysis. *Journal of Applied Psychology*, *93*(3), 540–567. https://doi.org/10.1037/0021-9010.93.3.540

Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of Job Performance Ratings: An Examination of Ratee Race, Ratee Gender, and Rater Level Effects. *Human Performance*, *9*(2), 103.

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology*, *74*(5), 770. https://doi.org/10.1037/0021-9010.74.5.770

Rahman, S., Hussain, B., & Haque, A. (2011). Organizational politics on employee performance: an exploratory study on readymade garments employees in Bangladesh. *Business Strategy Series*, *12*(3), 146–155. https://doi.org/10.1108/17515631111130112

Randall, R., & Sharples, D. (2012). The impact of rater agreeableness and rating context on the evaluation of poor performance. *Journal of Occupational and Organizational Psychology*, *85*(1), 42–59. https://doi.org/10.1348/2044-8325.002002

Reilly, R. R., Smither, J. W., & Vasilopoulos, N. L. (1996). A Longitudinal Study of Upward Feedback. *Personnel Psychology*, *49*(3), 599–612.

Renn, R. W., & Fedor, D. B. (2001). Development and field test of a feedback seeking, self-efficacy, and goal setting model of work performance. *Journal of Management*, *27*(5), 563–583. https://doi.org/10.1016/S0149-2063(01)00108-8

Robbins, T. L., & DeNisi, A. S. (1994). A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations. *Journal of Applied Psychology*, *79*(3), 341. https://doi.org/10.1037/0021-9010.79.3.341

Robbins, T. L., & DeNisi, A. S. (1998). Mood vs. Interpersonal Affect: Identifying Process and Rating Distortions in Performance Appraisal. *Journal of Business and Psychology*, *12*(3), 313–325.

Roch, S. G., Paquin, A. R., & Littlejohn, T. W. (2009). Do Raters Agree More on Observable Items? *Human Performance*, *22*(5), 391–409.

Rodgers, R., & Hunter, J. E. (1991). Impact of management by objectives on organizational productivity. *Journal of Applied Psychology*, *76*(2), 322. https://doi.org/10.1037/0021-9010.76.2.322

Rosch, D. M., Anderson, J. C., & Jordan, S. N. (2012). Analyzing the Effectiveness of Multisource Feedback as a Leadership Development Tool for College Students. *Journal of Leadership Studies*, *6*(3), 33–46. https://doi.org/10.1002/jls.21254

Rosen, B., & Jerdee, T. H. (1976b). The influence of age stereotypes on managerial decisions. *Journal of Applied Psychology*, *61*(4), 428. https://doi.org/10.1037/0021-9010.61.4.428

Rosen, B., & Jerdee, T. H. (1976a). The nature of job-related age stereotypes. *Journal of Applied Psychology*, *61*(2), 180. https://doi.org/10.1037/0021-9010.61.2.180

Rosen, C. C., Kacmar, K. M., Harris, K. J., Gavin, M. B., & Hochwarter, W. A. (2016). Workplace Politics and Performance Appraisal. *Journal of Leadership & Organizational Studies*. https://doi.org/10.1177/1548051816661480

Roth, P. L., Purvis, K. L., & Bobko, P. (2012). A Meta-Analysis of Gender Group Differences for Measures of Job Performance in Field Studies. *Journal of Management*. https://doi.org/10.1177/0149206310374774

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, *75*(3), 322. https://doi.org/10.1037/0021-9010.75.3.322

Rowson, A.-M. (1998). Using 360 Degree Feedback Instruments up, down and around the world: Implications for global implementation and use of Multi-Rater Feedback. *International Journal of Selection and Assessment*, *6*(1), 45–48. https://doi.org/10.1111/1468-2389.00071

Rupp, D. E., Hoffman, B. J., Bischof, D., Byham, W., Collins, L., Gibbons, A., … Thornton, G. (2015). Guidelines and Ethical Considerations for Assessment Center Operations. *Journal of Management*, *41*(4), 1244–1273. https://doi.org/10.1177/0149206314567780

Sackett, P. R., & DuBois, C. L. (1991). Rater-ratee race effects on performance evaluation: Challenging meta-analytic conclusions. *Journal of Applied Psychology*, *76*(6), 873. https://doi.org/10.1037/0021-9010.76.6.873

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review and critical reappraisal. *Psychological Bulletin*, *95*(3), 355. https://doi.org/10.1037/0033-2909.95.3.355

Salvemini, N. J., Reilly, R. R., & Smither, J. W. (1993). The Influence of Rater Motivation on Assimilation Effects and Accuracy in Performance Ratings. *Organizational Behavior and Human Decision Processes*, *55*(1), 41–60. https://doi.org/10.1006/obhd.1993.1023

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The triangle model of responsibility. *Psychological Review*, *101*(4), 632. https://doi.org/10.1037/0033-295X.101.4.632

Schmidt, F. L., & Hunter, J. (2004). General Mental Ability in the World of Work: Occupational Attainment and Job Performance. *Journal of Personality and Social Psychology*, *86*(1), 162. https://doi.org/10.1037/0022-3514.86.1.162

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, *71*(3), 432. https://doi.org/10.1037/0021-9010.71.3.432

Schmitt, N., & Lappin, M. (1980). Race and sex as determinants of the mean and variance of performance ratings. *Journal of Applied Psychology*, *65*(4), 428. https://doi.org/10.1037/0021-9010.65.4.428

Schneider, R. J., Goff, M., Anderson, S., & Borman, W. C. (2003). Computerized Adaptive Rating Scales for Measuring Managerial Performance. *International Journal of Selection & Assessment*, *11*(2/3), 237–246. https://doi.org/10.1111/1468-2389.00247

Schwab, D. P., & Heneman, H. G. (1978). Age stereotyping in performance appraisal. *Journal of Applied Psychology*, *63*(5), 573. https://doi.org/10.1037/0021-9010.63.5.573

Scullen, S. E., Mount, M. K., & Judge, T. A. (2003). Evidence of the Construct Validity of Developmental Ratings of Managerial Performance. *Journal of Applied Psychology*, *88*(1), 50–66. https://doi.org/10.1037/0021-9010.88.1.50

Seifert, C. F., Yukl, G., & McDonald, R. A. (2003). Effects of Multisource Feedback and a Feedback Facilitator on the Influence Behavior of Managers Toward Subordinates. *Journal of Applied Psychology*, *88*(3), 561–569. https://doi.org/10.1037/0021-9010.88.3.561

Shanock, L. R., Baran, B. E., Gentry, W. A., Pattison, S. C., & Heggestad, E. D. (2010). Polynomial Regression with Response Surface Analysis: A Powerful Approach for Examining Moderation and Overcoming Limitations of Difference Scores. *Journal of Business and Psychology*, *25*(4), 543–554. https://doi.org/10.1007/s10869-010-9183-4

Shore, L. M., & Thornton III, G. C. (1986). Effects of Gender on Self- and Supervisory Ratings. *Academy of Management Journal*, *29*(1), 115–129. https://doi.org/10.2307/255863

Shore, T. H., & Tashchain, A. (2002). Accountability Forces in Performance Appraisal: Effects of Self-Appraisal Information, Normative Information, and Task Performance. *Journal of Business and Psychology*, *17*(2), 261–274.

Smith, D. E. (1986). Training Programs for Performance Appraisal: A Review. *Academy of Management Review*, *11*(1), 22–40. https://doi.org/10.5465/AMR.1986.4282615

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, *47*(2), 149. https://doi.org/10.1037/h0047060

Smith, W. J., Harrington, K. V., & Houghton, J. D. (2000). Predictors of performance appraisal discomfort. *Public Personnel Management*, *29*(1), 21–21.

Smither, J. W. (2012). Performance Management. In S. W. J. Kozlowski (Ed.), *The Oxford Handbook of Organizational Psychology, Volume 1*. Oxford University Press. Retrieved from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199928309.001.0001/oxfordhb-9780199928309-e-10

Smither, J. W., London, M., Flautt, R., Vargas, Y., & Kucine, I. (2003). Can Working with an Executive Coach Improve Multisource Feedback Ratings Over Time? a Quasi-Experimental Field Study. *Personnel Psychology*, *56*(1), 23–44.

Smither, J. W., London, M., & Reilly, R. R. (2005). Does Performance Improve Following Multisource Feedback? a Theoretical Model, Meta-Analysis, and Review of Empirical Findings. *Personnel Psychology*, *58*(1), 33–66. https://doi.org/10.1111/j.1744-6570.2005.514_1.x

Smither, J. W., London, M., & Richmond, K. R. (2005). The Relationship Between Leaders' Personality and Their Reactions to and Use of Multisource Feedback A

Longitudinal Study. *Group & Organization Management*, *30*(2), 181–210. https://doi.org/10.1177/1059601103254912

Smither, J. W., London, M., Vasilopoulos, N. L., Reilly, R. R., Millsap, R. E., & Salvemini, N. (1995). An Examination of the Effects of an Upward Feedback Program Over Time. *Personnel Psychology*, *48*(1), 1–34.

Smither, J. W., & Reilly, R. R. (1987). True intercorrelation among job components, time delay in rating, and rater intelligence as determinants of accuracy in performance ratings. *Organizational Behavior and Human Decision Processes*, *40*(3), 369–391. https://doi.org/10.1016/0749-5978(87)90022-7

Smither, J. W., Reilly, R. R., & Buda, R. (1988). Effect of prior performance information on ratings of present performance: Contrast versus assimilation revisited. *Journal of Applied Psychology*, *73*(3), 487. https://doi.org/10.1037/0021-9010.73.3.487

Smither, J. W., & Walker, A. G. (2004). Are the Characteristics of Narrative Comments Related to Improvement in Multirater Feedback Ratings Over Time? *Journal of Applied Psychology*, *89*(3), 575–581. https://doi.org/10.1037/0021-9010.89.3.575

Smither, J. W., Walker, A. G., & Yap, M. K. T. (2004). An Examination of the Equivalence of Web-Based Versus Paper-and-Pencil Upward Feedback Ratings: Rater- and Ratee-Level Analyses. *Educational and Psychological Measurement*, *64*(1), 40–61. https://doi.org/10.1177/0013164403258429

Sosik, J. J. (2001). Self-Other Agreement on Charismatic Leadership Relationships with Work Attitudes and Managerial Performance. *Group & Organization Management*, *26*(4), 484–511. https://doi.org/10.1177/1059601101264005

Sosik, J. J., & Godshalk, V. M. (2004). Self-Other Rating Agreement in Mentoring Meeting Protégé Expectations for Development and Career Advancement. *Group & Organization Management*, *29*(4), 442–469. https://doi.org/10.1177/1059601103257421

Sosik, J. J., & Megerian, L. E. (1999). Understanding Leader Emotional Intelligence and Performance The Role of Self-Other Agreement on Transformational Leadership Perceptions. *Group & Organization Management*, *24*(3), 367–390. https://doi.org/10.1177/1059601199243006

Spence, J. R., & Keeping, L. (2011). Conscious rating distortion in performance appraisal: A review, commentary, and proposed framework for research. *Human Resource Management Review*, *21*(2), 85–95. https://doi.org/10.1016/j.hrmr.2010.09.013

Spence, J. R., & Keeping, L. M. (2010). The impact of non-performance information on ratings of job performance: A policy-capturing approach. *Journal of Organizational Behavior*, *31*(4), 587–608.

Spence, J. R., & Keeping, L. M. (2013). The road to performance ratings is paved with intentions A framework for understanding managers' intentions when rating employee performance. *Organizational Psychology Review*, *3*(4), 360–383. https://doi.org/10.1177/2041386613485969

Spool, M. D. (1978). Training Programs for Observers of Behavior: A Review. *Personnel Psychology*, *31*(4), 853–888.

Steel, R. P., & Ovalle, N. K. (1984). Self-Appraisal Based Upon Supervisory Feedback. *Personnel Psychology*, *37*(4), 667–685. https://doi.org/10.1111/j.1744-6570.1984.tb00532.x

Strauss, J. P., Barrick, M. R., & Connerley, M. L. (2001). An investigation of personality similarity effects (relational and perceived) on peer and supervisor ratings and the role of familiarity and liking. *Journal of Occupational and Organizational Psychology*, *74*(5), 637–657. https://doi.org/10.1348/096317901167569

Stroh, L. K., Brett, J. M., & Reilly, A. H. (1992). All the right stuff: A comparison of female and male managers' career progression. *Journal of Applied Psychology*, *77*(3), 251–260. https://doi.org/10.1037/0021-9010.77.3.251

Sutton, A. W., Baldwin, S. P., Wood, L., & Hoffman, B. J. (2013). A Meta-Analysis of the Relationship Between Rater Liking and Performance Ratings. *Human Performance*, *26*(5), 409–429.

Szell, S., & Henderson, R. (1997). The impact of self-supervisor/subordinate performance rating agreement on subordinates' job satisfaction and organisational commitment. *Journal of Applied Social Behaviour*, *3*(2), 25–37.

Taft, R. (1955). The ability to judge people. *Psychological Bulletin*, *52*(1), 1. https://doi.org/10.1037/h0044999

Tang, K. Y., Dai, G., & Meuse, K. P. D. (2013). Assessing leadership derailment factors in 360° feedback: Differences across position levels and self-other agreement. *Leadership & Organization Development Journal*, *34*(4), 326–343. https://doi.org/10.1108/LODJ-07-2011-0070

Taylor, E. K., & Wherry, R. J. (1951). A Study of Leniency in Two Rating Systems. *Personnel Psychology*, *4*(1), 39–47. https://doi.org/10.1111/j.1744-6570.1951.tb01459.x

Taylor, M. S., Fisher, C. D., & Ilgen, D. R. (1984). Individuals' Reactions to Performance Feedback in Organizations: A Control Theory Perspective. *Research in Personnel and Human Resources*, *2*, 81–124.

Taylor, S. N., & Bright, D. S. (2011). Open-Mindedness and Defensiveness in Multisource Feedback Processes A Conceptual Framework. *The Journal of Applied Behavioral Science*, *47*(4), 432–460. https://doi.org/10.1177/0021886311408724

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality Measures as Predictors of Job Performance: A Meta-Analytic Review. *Personnel Psychology*, *44*(4), 703–742.

Thach, E. C. (2002). The impact of executive coaching and 360 feedback on leadership effectiveness. *Leadership & Organization Development Journal*, *23*(4), 205–214. https://doi.org/10.1108/01437730210429070

Thorndike, R. L. (1949). *Personnel Selection: Test and Measurement Techniques*. New York: Wiley & Sons, Inc.

Thorsteinson, T. J., Breier, J., Atwell, A., Hamilton, C., & Privette, M. (2008). Anchoring effects on performance judgments. *Organizational Behavior and Human Decision Processes*, *107*(1), 29–40. https://doi.org/10.1016/j.obhdp.2008.01.003

Tornow, W. W. (1993). Perceptions or reality: Is multi-perspective measurement a means or an end? *Human Resource Management*, *32*(2–3), 221–229. https://doi.org/10.1002/hrm.3930320203

Tsui, A. S., & Barry, B. (1986). Research Notes. Interpersonal Affect and Rating Errors. *Academy of Management Journal*, *29*(3), 586–599. https://doi.org/10.2307/256225

Tsui, A. S., & O'Reilly III, C. A. (1989). Beyond Simple Demographic Effects: The Importance of Relational Demography in Superior-Subordinate Dyads. *Academy of Management Journal*, *32*(2), 402–423. https://doi.org/10.2307/256368

Turban, D. B., Jones, A. P., & Rozelle, R. M. (1990). Influences of supervisor liking of a subordinate and the reward context on the treatment and evaluation of that subordinate. *Motivation and Emotion*, *14*(3), 215–233. https://doi.org/10.1007/BF00995570

Tziner, A. (1999). The Relationship Between Distal and Proximal Factors and the Use of Political Considerations in Performance Appraisal. *Journal of Business and Psychology*, *14*(1), 217–231. https://doi.org/10.1023/A:1022931106379

Tziner, A., Joanis, C., & Murphy, K. R. (2000). A Comparison of Three Methods of Performance Appraisal with Regard to Goal Properties, Goal Perception, and Ratee Satisfaction. *Group & Organization Management*, *25*(2), 175–190. https://doi.org/10.1177/1059601100252005

Tziner, A., & Kopelman, R. (1988). Effects of rating format on goal-setting dimensions: A field experiment. *Journal of Applied Psychology*, *73*(2), 323. https://doi.org/10.1037/0021-9010.73.2.323

Tziner, A., Kopelman, R., & Joanis, C. (1997). Investigation of Raters' and Ratees' Reactions to Three Methods of Performance Appraisal: BOS, BARS, and GRS. *Canadian Journal of Administrative Sciences / Revue Canadienne Des Sciences de l'Administration*, *14*(4), 396–404. https://doi.org/10.1111/j.1936-4490.1997.tb00145.x

Tziner, A., Latham, G. P., Price, B. S., & Haccoun, R. (1996). Development and Validation of a Questionnaire for Measuring Perceived Political Considerations in Performance Appraisal. *Journal of Organizational Behavior*, *17*(2), 179–190.

Tziner, A., Murphy, K., Cleveland, J. N., Yavo, A., & Hayoon, E. (2008). A New Old Question: Do contextual factors relate to rating behavior: An investigation with peer evaluations. *International Journal of Selection and Assessment*, *16*(1), 59–67. https://doi.org/10.1111/j.1468-2389.2008.00409.x

Tziner, A., & Murphy, K. R. (1999). Additional Evidence of Attitudinal Influences in Performance Appraisal. *Journal of Business and Psychology*, *13*(3), 407–419. https://doi.org/10.1023/A:1022982501606

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2005). Contextual and Rater Factors Affecting Rating Behavior. *Group & Organization Management*, *30*(1), 89–98. https://doi.org/10.1177/1059601104267920

Tziner, A., Murphy, K. R., Cleveland, J. N., Beaudin, G., & Marchand, S. (1998). Impact of Rater Beliefs Regarding Performance Appraisal and Its Organizational Context on Appraisal Quality. *Journal of Business and Psychology*, *12*(4), 457–467. https://doi.org/10.1023/A:1025003106150

Tziner, A., Murphy, K. R., Cleveland, J. N., & Roberts-Thompson, G. P. (2001). Relationships Between Attitudes Toward Organizations and Performance Appraisal Systems and Rating Behavior. *International Journal of Selection & Assessment*, *9*(3), 226.

Tziner, A., Prince, J. B., & Murphy, K. R. (1997). PCPAQ -- The Questionnaire for Measuring Perceived Political Considerations in Performance Appraisal: Some New Evidence Regarding its Psychometric Qualities. *Journal of Social Behavior & Personality*, *12*(1), 189–199.

U.S. Department of Labor, Bureau of Labor Statistics. (2016). *Employee Tenure in 2016* [Press Release]. Retrieved from: https://www.bls.gov/news.release/pdf/tenure.pdf

van der Heijden, B. I. J. M., & Nijhof, A. H. J. (2004). The value of subjectivity: problems and prospects for 360-degree appraisal systems. *International Journal of Human Resource Management*, *15*(3), 493–511. https://doi.org/10.1080/0958519042000181223

Van Dijk, D., & Kluger, A. N. (2011). Task type as a moderator of positive/negative feedback effects on motivation and performance: A regulatory focus perspective. *Journal of Organizational Behavior*, *32*(8), 1084–1105. https://doi.org/10.1002/job.725

Van Velsor, E., Taylor, S., & Leslie, J. B. (1993). An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness. *Human Resource Management*, *32*(2–3), 249–263. https://doi.org/10.1002/hrm.3930320205

Vancouver, J. B., & Tischner, E. C. (2004). The Effect of Feedback Sign on Task Performance Depends on Self-Concept Discrepancies. *Journal of Applied Psychology*, *89*(6), 1092–1098. https://doi.org/10.1037/0021-9010.89.6.1092

Van-Dijk, D., & Kluger, A. N. (2004). Feedback Sign Effect on Motivation: Is it Moderated by Regulatory Focus? *Applied Psychology*, *53*(1), 113–135. https://doi.org/10.1111/j.1464-0597.2004.00163.x

Vanneste, B. S., Puranam, P., & Kretschmer, T. (2014). Trust over time in exchange relationships: Meta-analysis and theory. *Strategic Management Journal*, *35*(12), 1891–1902. https://doi.org/10.1002/smj.2198

Varma, A., Denisi, A. S., & Peters, L. H. (1996). Interpersonal Affect and Performance Appraisal: A Field Study. *Personnel Psychology*, *49*(2), 341–360.

Varma, A., & Pichler, S. (2007). Interpersonal Affect: Does It Really Bias Performance Appraisals? *Journal of Labor Research*, *28*(2), 397–412.

Varma, A., Pichler, S., & Srinivas, E. S. (2005). The role of interpersonal affect in performance appraisal: evidence from two samples – the US and India. *International Journal of Human Resource Management*, *16*(11), 2029–2044. https://doi.org/10.1080/09585190500314904

Vecchio, R. P. (1993). The impact of differences in subordinate and supervisor age on attitudes and performance. *Psychology and Aging*, *8*(1), 112. https://doi.org/10.1037/0882-7974.8.1.112

Vecchio, R. P. (1998). Leader-Member Exchange, Objective Performance, Employment Duration, and Supervisor Ratings: Testing for Moderation and Mediation. *Journal of Business and Psychology*, *12*(3), 327–341.

Vecchio, R. P., & Anderson, R. J. (2009). Agreement in Self–Other Ratings of Leader Effectiveness: The role of demographics and personality. *International Journal of Selection and Assessment*, *17*(2), 165–179. https://doi.org/10.1111/j.1468-2389.2009.00460.x

Vigoda E. (2000). Internal politics in public administration systems. *Public Personnel Management*, *29*(2), 185–185.

Villanova, P., & Bernardin, H. J. (1989). Impression Management in the Context of Performance Appraisal. In R. A. Giacalone & P. Rosenfeld (Eds.), *Impression Management in the Organization* (pp. 299–313). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Villanova, P., Bernardin, H. J., Dahmus, S. A., & Sims, R. L. (1993). Rater Leniency and Performance Appraisal Discomfort. *Educational and Psychological Measurement*, *53*(3), 789–799. https://doi.org/10.1177/0013164493053003023

Villanova, P., Bernardin, H. J., & Ross, S. (1997). *The prediction of rater leniency using individual difference measures.* Unpublished Manuscript.

Violato, C., Lockyer, J. M., & Fidler, H. (2008). Changes in performance: a 5-year longitudinal study of participants in a multi-source feedback programme. *Medical Education*, *42*(10), 1007–1013. https://doi.org/10.1111/j.1365-2923.2008.03127.x

Visser, B. A., Ashton, M. C., & Vernon, P. A. (2008). What makes you think you're so smart? Measured abilities, personality, and sex differences in relation to self-estimates of multiple intelligences. *Journal of Individual Differences*, *29*(1), 35–44. https://doi.org/10.1027/1614-0001.29.1.35

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*(5), 557–574. https://doi.org/10.1037/0021-9010.81.5.557

Visweswaran, C. (2001). Assessment of individual job performance: A review of the past century and a look ahead. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Visweswaran (Eds.), *Handbook of industrial, work, and organizational psychology* (Vol. 1, pp. 110–126). Thousand Oaks, CA: Sage.

Waldman, D. A., & Avolio, B. J. (1991). Race effects in performance evaluations: Controlling for ability, education, and experience. *Journal of Applied Psychology*, *76*(6), 897. https://doi.org/10.1037/0021-9010.76.6.897

Waldman, D., Atwater, L. E., & Antonioni, D. (1998). Has 360 Degree Feedback Gone Amok? *The Academy of Management Executive*, *12*(2), 86–94.

Walker, A. G., & Smither, J. W. (1999). A Five-Year Study of Upward Feedback: What Managers Do with Their Results Matters. *Personnel Psychology*, *52*(2), 393–423.

Walker, A. G., Smither, J. W., Atwater, L. E., Dominick, P. G., Brett, J. F., & Reilly, R. R. (2010). Personality and Multisource Feedback Improvement: A Longitudinal Investigation. *Journal of Behavioral & Applied Management*, *11*(2), 175–204.

Wallace, L. E., Stelman, S. A., & Chaffee, D. S. (2016). Ratee Reactions Drive Performance Appraisal Success (and Failure). *Industrial and Organizational Psychology*, *9*(2), 310–314. https://doi.org/10.1017/iop.2016.16

Wanous, J. P., & Hudy, M. J. (2001). Single-Item Reliability: A Replication and Extension. *Organizational Research Methods*, *4*(4), 361–375. https://doi.org/10.1177/109442810144003

Warech, M. A., Smither, J. W., Reilly, R. R., Millsap, R. E., & Reilly, S. P. (1998). Self-monitoring and 360-degree ratings. *The Leadership Quarterly*, *9*(4), 449–473. https://doi.org/10.1016/S1048-9843(98)90011-X

Waung, M., & Highhouse, S. (1997). Fear of Conflict and Empathic Buffering: Two Explanations for the Inflation of Performance Feedback. *Organizational Behavior and Human Decision Processes*, *71*(1), 37–54. https://doi.org/10.1006/obhd.1997.2711

Wayne, S. J., & Ferris, G. R. (1990). Influence tactics, affect, and exchange quality in supervisor-subordinate interactions: A laboratory experiment and field study. *Journal of Applied Psychology*, *75*(5), 487. https://doi.org/10.1037/0021-9010.75.5.487

Weisband, S., & Atwater, L. (1999). Evaluating self and others in electronic and face-to-face groups. *Journal of Applied Psychology*, *84*(4), 632. https://doi.org/10.1037/0021-9010.84.4.632

Weiss, D. J. (2011). Better Data From Better Measurements Using Computerized Adaptive Testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1–27.

Wesman, A. G. (1965). *Wesman personnel classification test*. Psychological Corporation.

Wexley, K. N., & Pulakos, E. D. (1982). Sex effects on performance ratings on manager–subordinate dyads: A field study. *Journal of Applied Psychology*, *67*(4), 433. https://doi.org/10.1037/0021-9010.67.4.433

Wexley, K. N., & Youtz, M. A. (1985). Rater beliefs about others: Their effects on rating errors and rater accuracy. *Journal of Occupational Psychology*, *58*(4), 265–275.

Whitaker, B. G., Dahling, J. J., & Levy, P. (2007). The Development of a Feedback Environment and Role Clarity Model of Job Performance. *Journal of Management*, *33*(4), 570–591. https://doi.org/10.1177/0149206306297581

Williams, J. R., Miller, C. E., Steelman, L. A., & Levy, P. E. (1999). Increasing feedback seeking in public contexts: It takes two (or more) to tango. *Journal of Applied Psychology*, *84*(6), 969–976. https://doi.org/10.1037/0021-9010.84.6.969

Wilson, K. Y. (2010). An analysis of bias in supervisor narrative comments in performance appraisal. *Human Relations*, *63*(12), 1903–1933. https://doi.org/10.1177/0018726710369396

Wimer, S., & Nowack, K. M. (1998). 13 Common Mistakes Using 360-Degree Feedback. *Training & Development*, *52*, 69–82.

Witt, L. A. (1998). Enhancing organizational goal congruence: A solution to organizational politics. *Journal of Applied Psychology*, *83*(4), 666. https://doi.org/10.1037/0021-9010.83.4.666

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational & Organizational Psychology*, *67*(3), 189–205.

Woehr, D. J., & Roch, S. G. (2016). Of Babies and Bathwater: Don't Throw the Measure Out With the Application. *Industrial and Organizational Psychology*, *9*(2), 357–361. https://doi.org/10.1017/iop.2016.25

Wohlers, A. I., & London, M. (1989). Ratings of Managerial Characteristics: Evaluation Difficulty, Co-Worker Agreement, and Self-Awareness. *Personnel Psychology*, *42*(2), 235–261. https://doi.org/10.1111/j.1744-6570.1989.tb00656.x

Woo, S. E., Sims, C. S., Rupp, D. E., & Gibbons, A. M. (2008). Development Engagement Within and Following Developmental Assessment Centers: Considering Feedback Favorability and Self–Assessor Agreement. *Personnel Psychology*, *61*(4), 727–759. https://doi.org/10.1111/j.1744-6570.2008.00129.x

Yammarino, F. J. (1998). Multivariate aspects of the Varient/WABA approach: A discussion and leadership illustration. *Leadership Quarterly*, *9*(2), 203.

Yammarino, F. J., & Atwater, L. E. (1993). Understanding self-perception accuracy: Implications for human resource management. *Human Resource Management*, *32*(2–3), 231–247. https://doi.org/10.1002/hrm.3930320204

Yammarino, F. J., & Atwater, L. E. (1997). Do Managers See Themselves as Others See Them? Implications of Self-Other Rating Agreement for Human Resources Management. *Organizational Dynamics*, *25*(4), 35–44.

Yukl, G., Gordon, A., & Taber, T. (2002). A Hierarchical Taxonomy of Leadership Behavior: Integrating a Half Century of Behavior Research. *Journal of Leadership & Organizational Studies*, *9*(1), 15–32. https://doi.org/10.1177/107179190200900102

Yukl, G., & Lepsinger, R. (1995, December). How to get the most out of 360 degree feedback. *Training*, *32*(12), 45+.

Yun, G. J., Donahue, L. M., Dudley, N. M., & McFarland, L. A. (2005). Rater Personality, Rating Format, and Social Context: Implications for Performance Appraisal Ratings. *International Journal of Selection & Assessment*, *13*(2), 97–107. https://doi.org/10.1111/j.0965-075X.2005.00304.x

Zalesny, M. D., & Kirsch, M. P. (1989). The Effect of Similarity on Performance Ratings and Interrater Agreement. *Human Relations, 42*(1), 81–96. https://doi.org/10.1177/001872678904200105

Zedeck, S., & Baker, H. T. (1972). Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance, 7*(3), 457–466. https://doi.org/10.1016/0030-5073(72)90029-3

Zheng, X., Diaz, I., Jing, Y., & Chiaburu, D. S. (2015). Positive and negative supervisor developmental feedback and task-performance. *Leadership & Organization Development Journal*, *36*(2), 212–232. https://doi.org/10.1108/LODJ-04-2013-0039

Zivnuska, S., Kacmar, K. M., Witt, L. A., Carlson, D. S., & Bratton, V. K. (2004). Interactive Effects of Impression Management and Organizational Politics on Job Performance. *Journal of Organizational Behavior*, *25*(5), 627–640.

Zyphur, M. J., Chaturvedi, S., & Arvey, R. D. (2008). Job performance over time is a function of latent trajectories and previous performance. *Journal of Applied Psychology*, *93*(1), 217. https://doi.org/10.1037/0021-9010.93.1.217

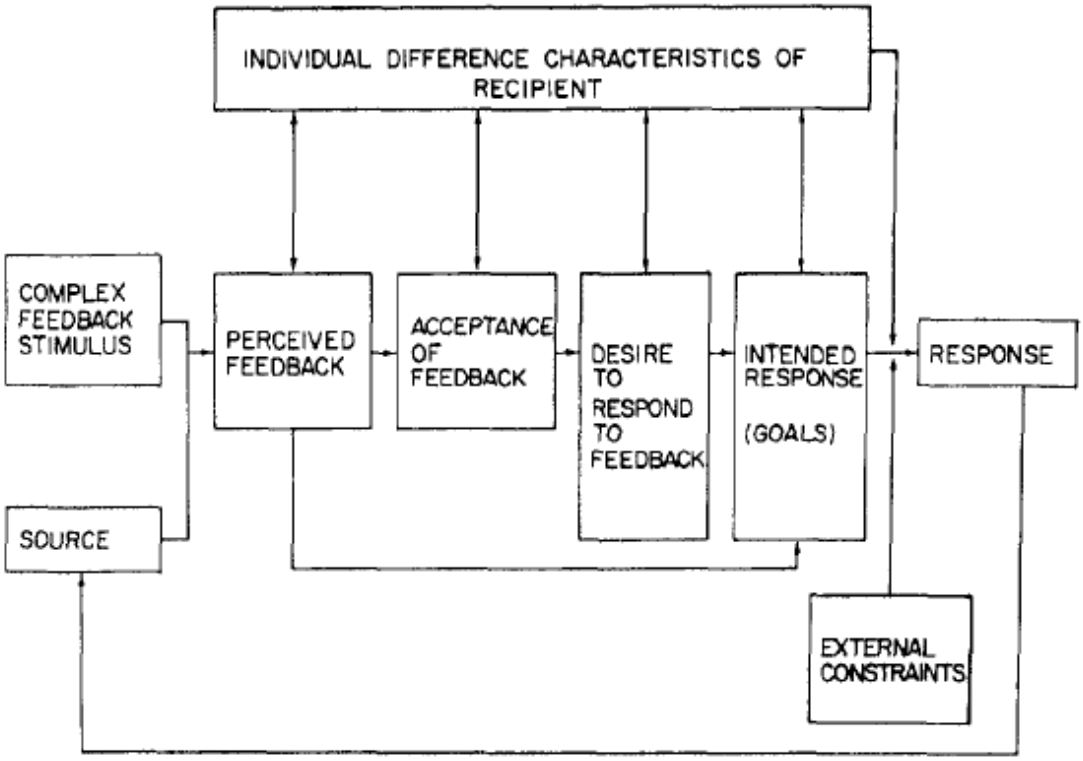

Figure 1: Ilgen, Fisher, and Taylor's (1979) model of the effects of feedback on recipients.

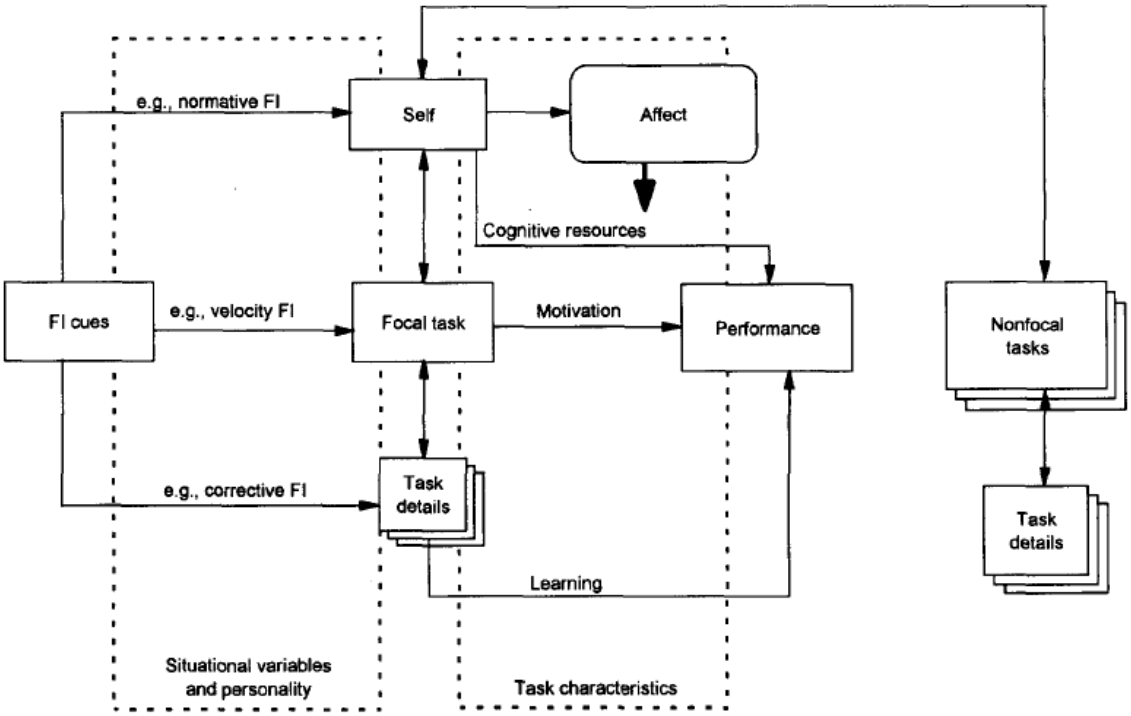

Figure 2: An overview of Kluger and DeNisi's (1996) Feedback Intervention Theory (FIT).

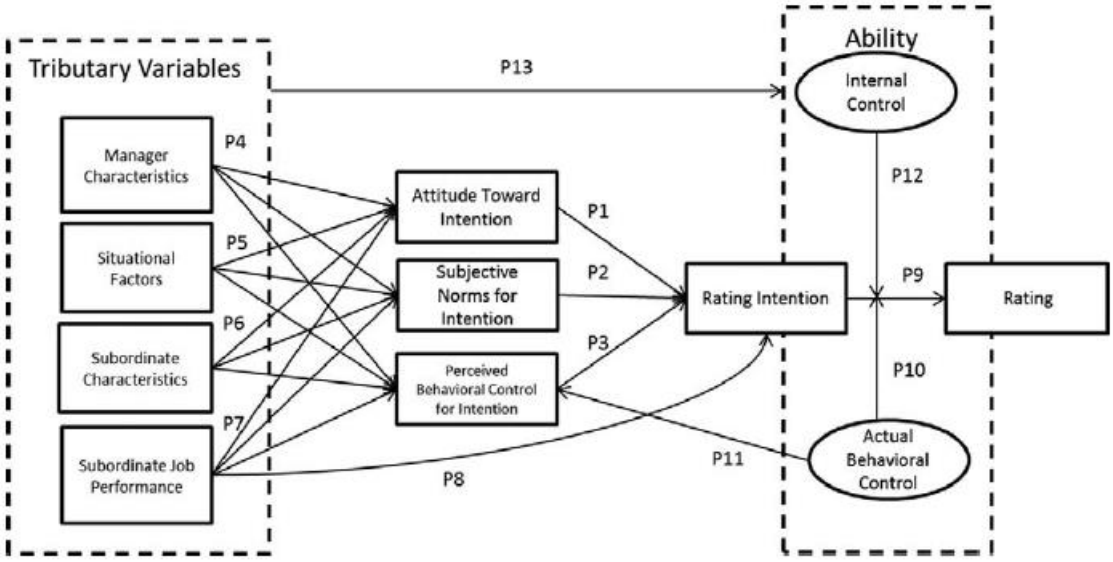

Figure 3: An overview of Spence & Keeping's (2013) model of motivation of job performance raters.

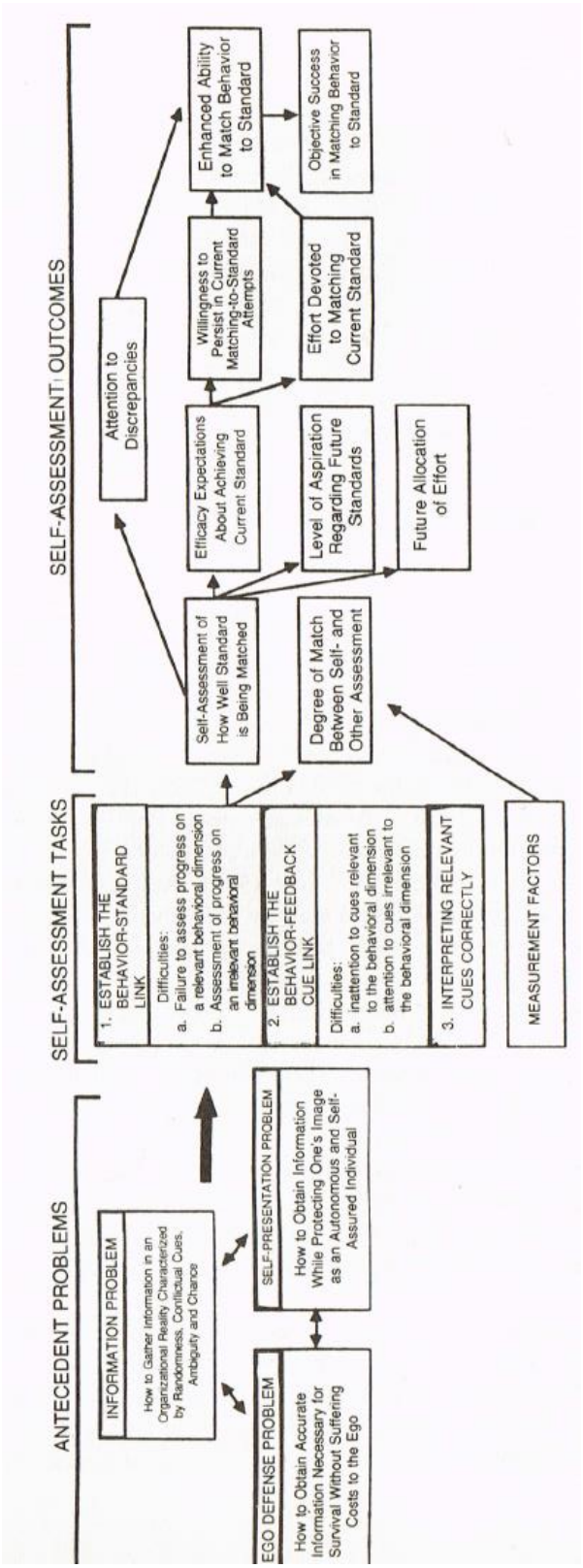

Figure 4: Overview of Ashford's (1989) model of self-assessment.

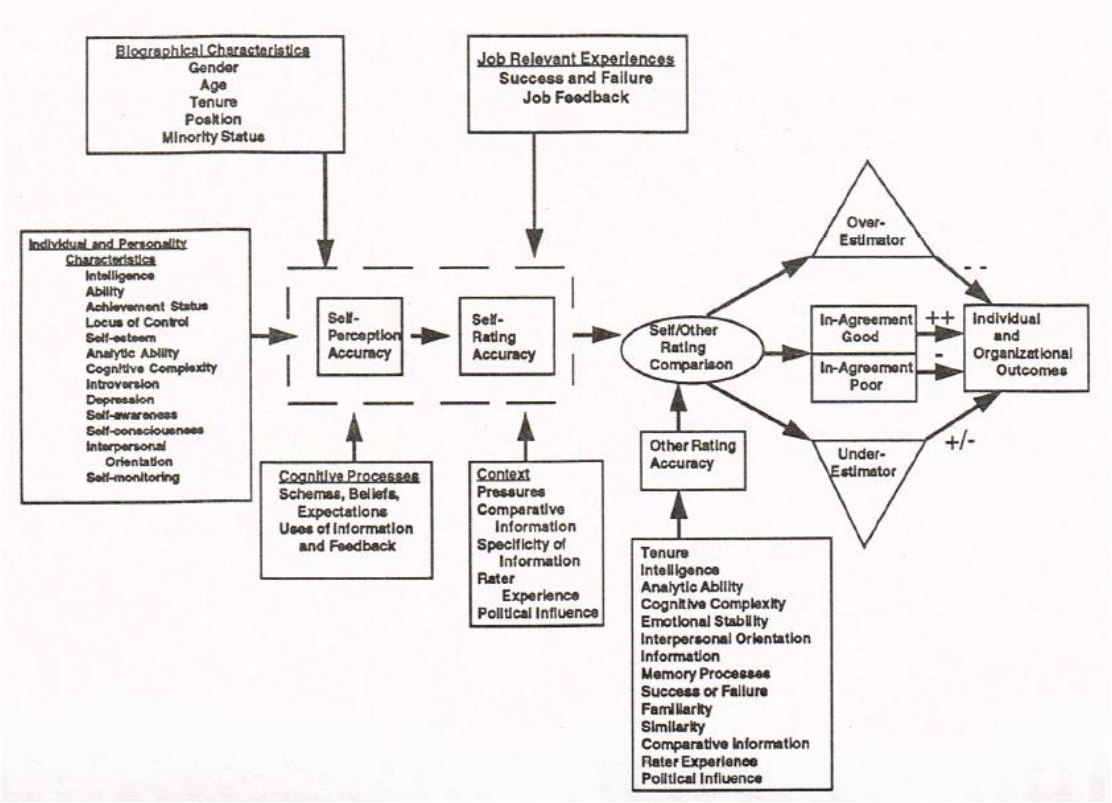

Figure 5: Overview of Atwater & Yammarino's (1997) model of self-other agreement in job performance ratings.

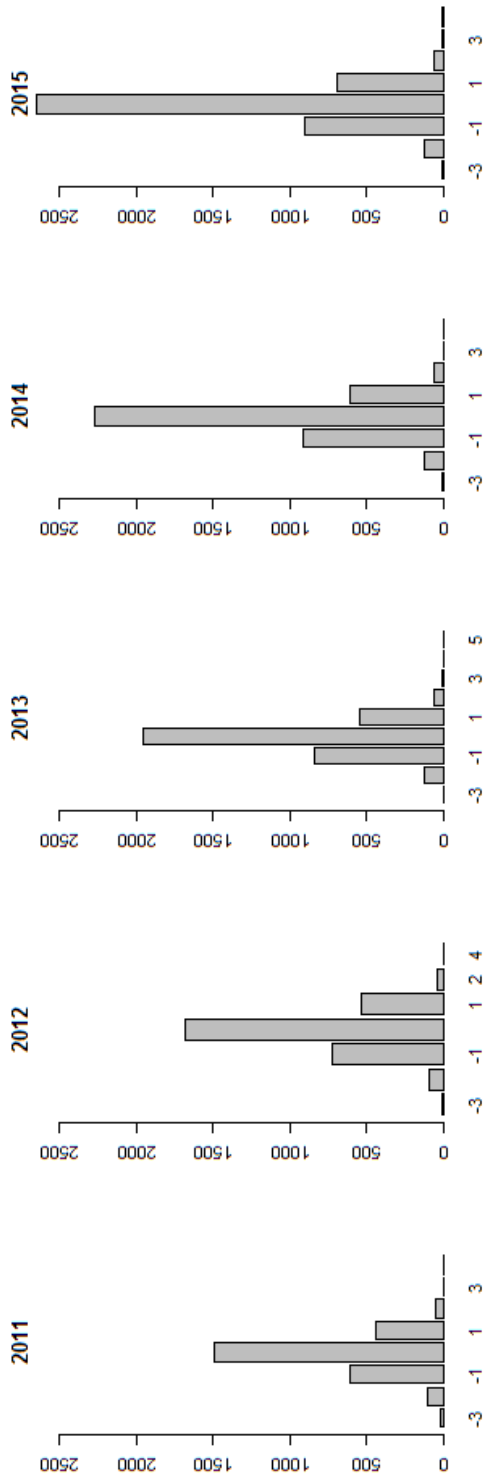

Figure 6: Histograms for each year's differences in performance appraisal ratings.

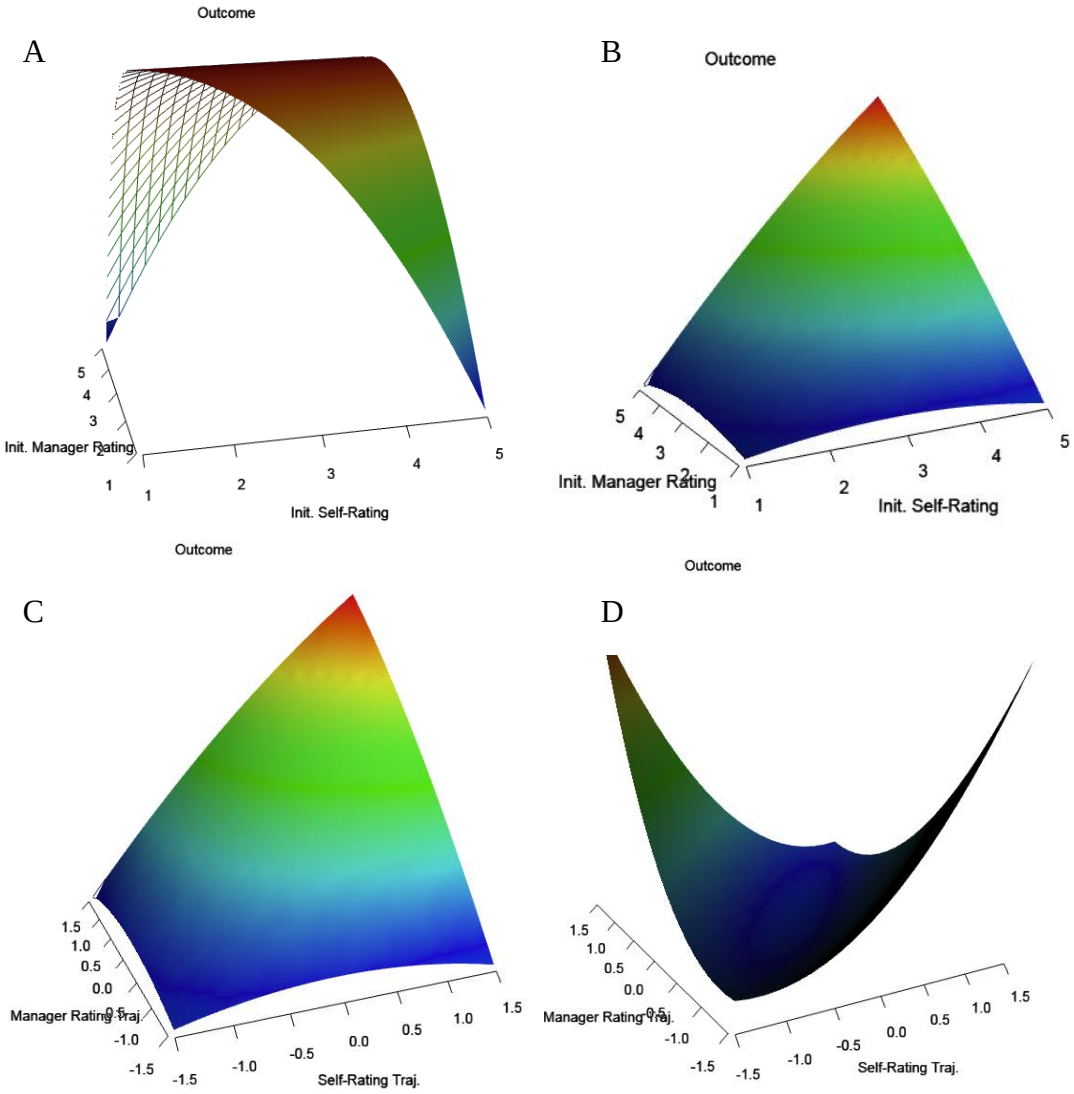

Figure 7: Hypothetical graphs to illustrate potential outcomes of polynomial analyses.

For intercept effects: A) Agreement is unilaterally better for outcomes, regardless of performance level. B) Agreement's effects depend on the level of performance. For slope effects: C) Agreement on growth is better than agreement on decline. D) Convergence over time is best.

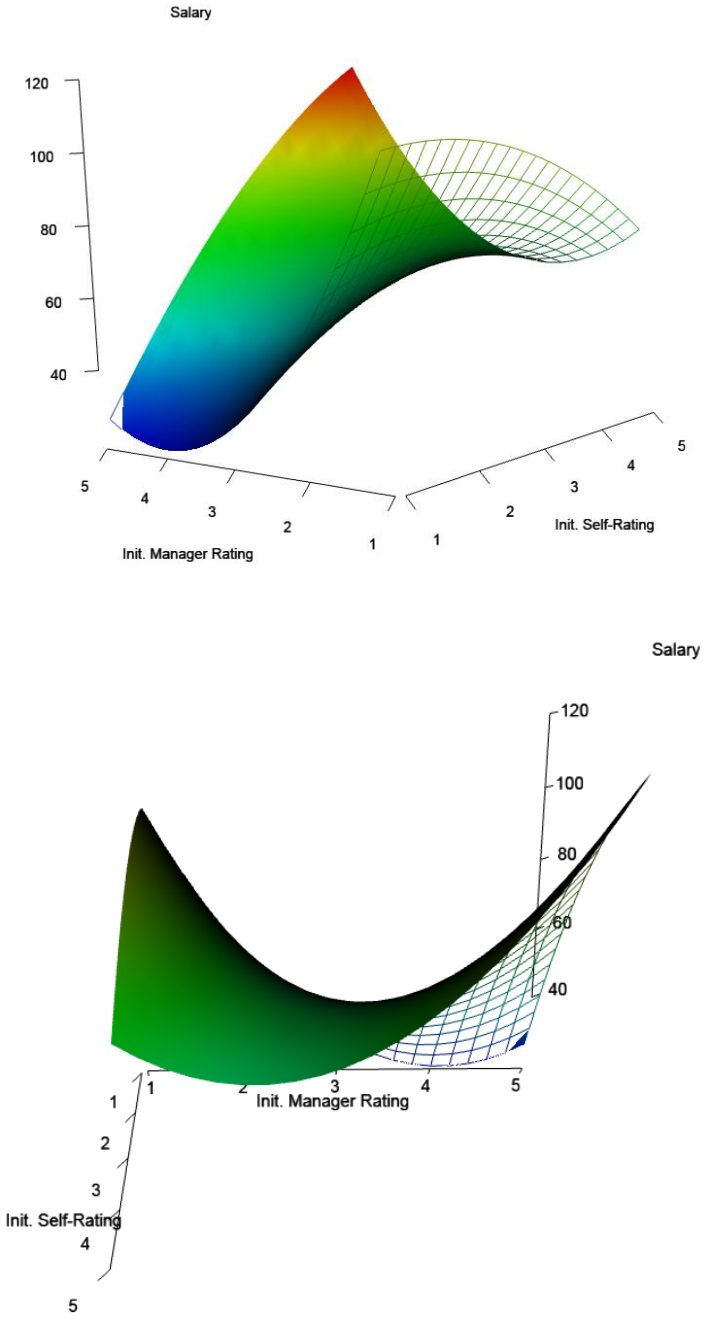

Figure 8: Response surface for initial salary (in $1k) predicted by initial performance ratings.

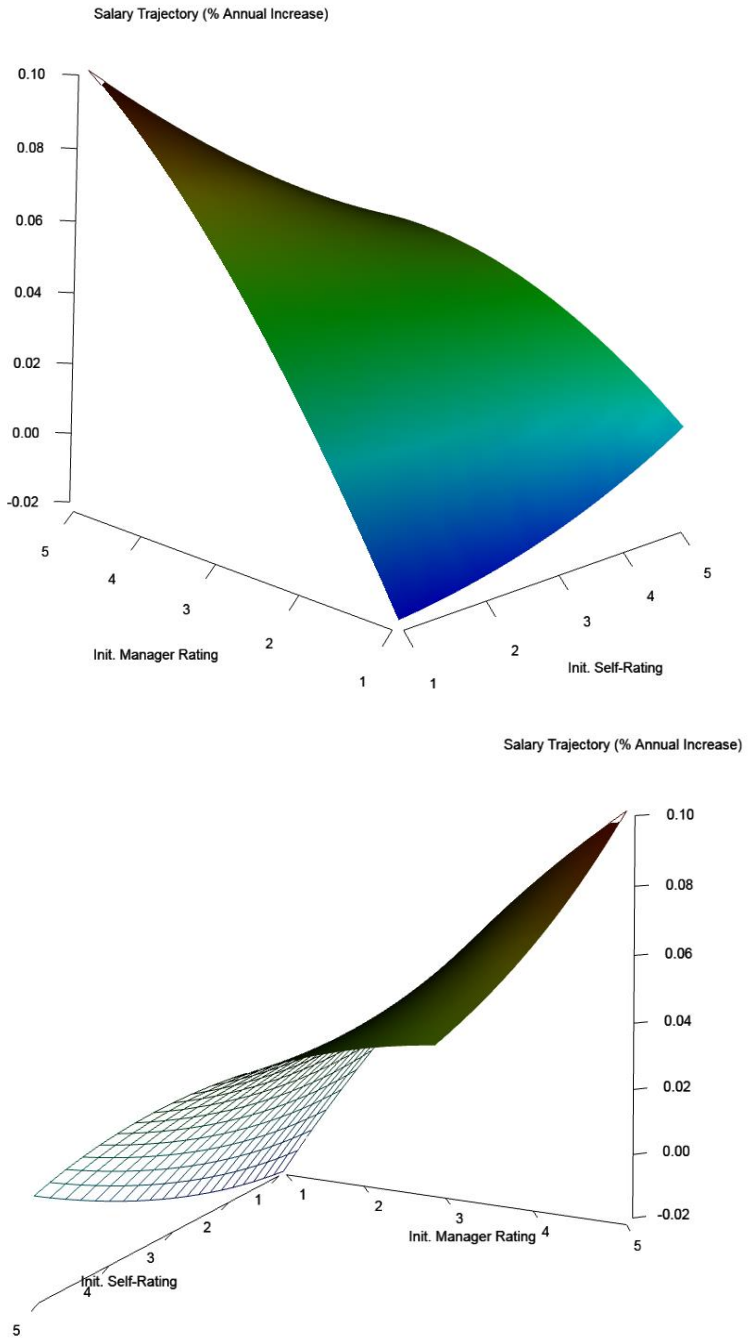

Figure 9: Response surface for salary trajectories predicted by initial ratings.

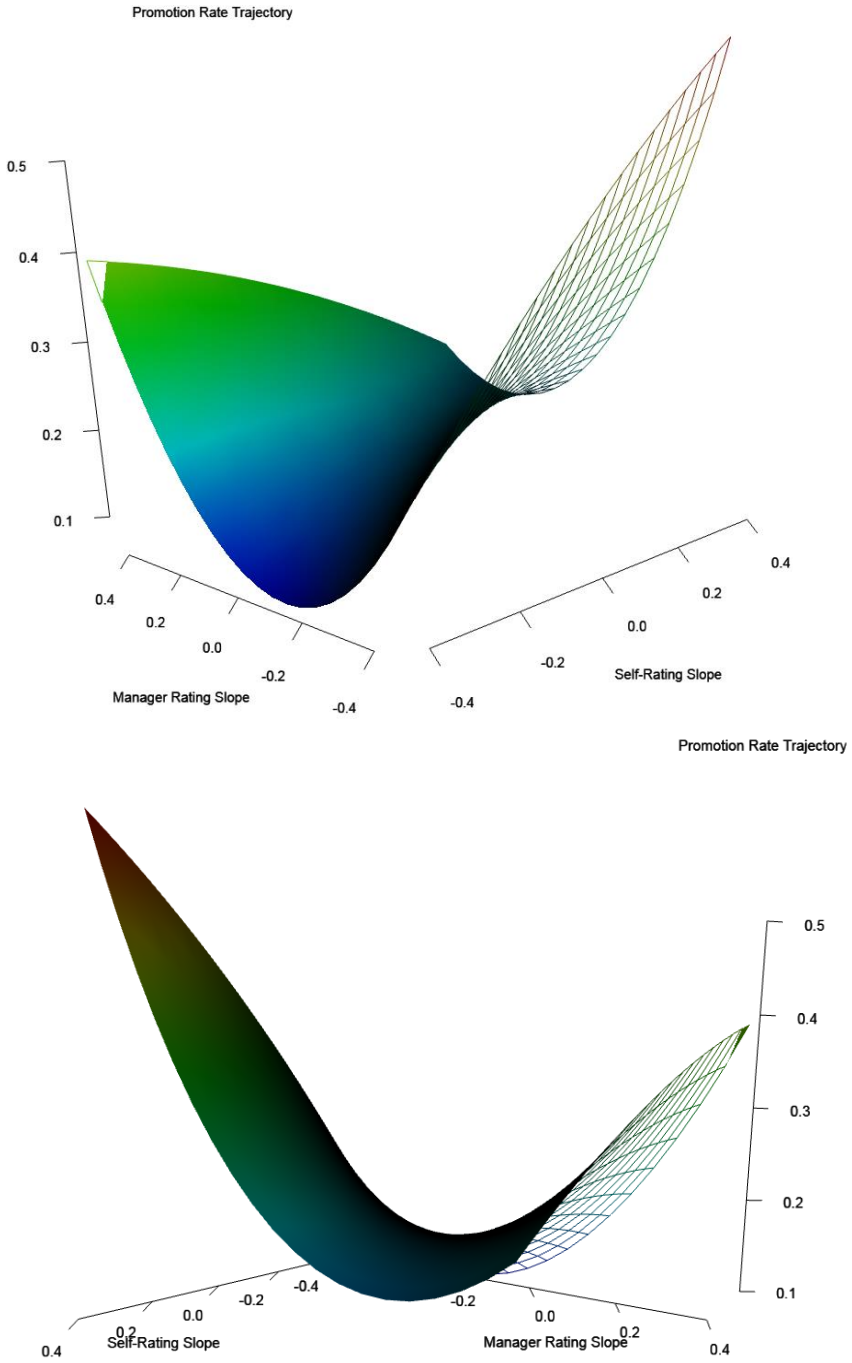

Figure 10: Response surface for promotion trajectories (expected # promotions per year) predicted by rating trajectories.

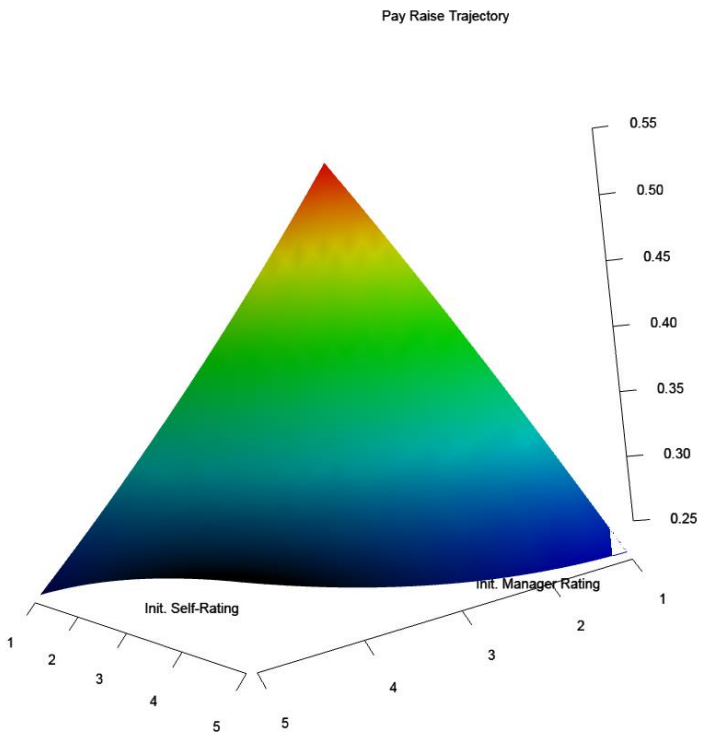

Figure 11: Response surface for pay raise trajectories (in # raises per year) predicted by initial performance ratings.

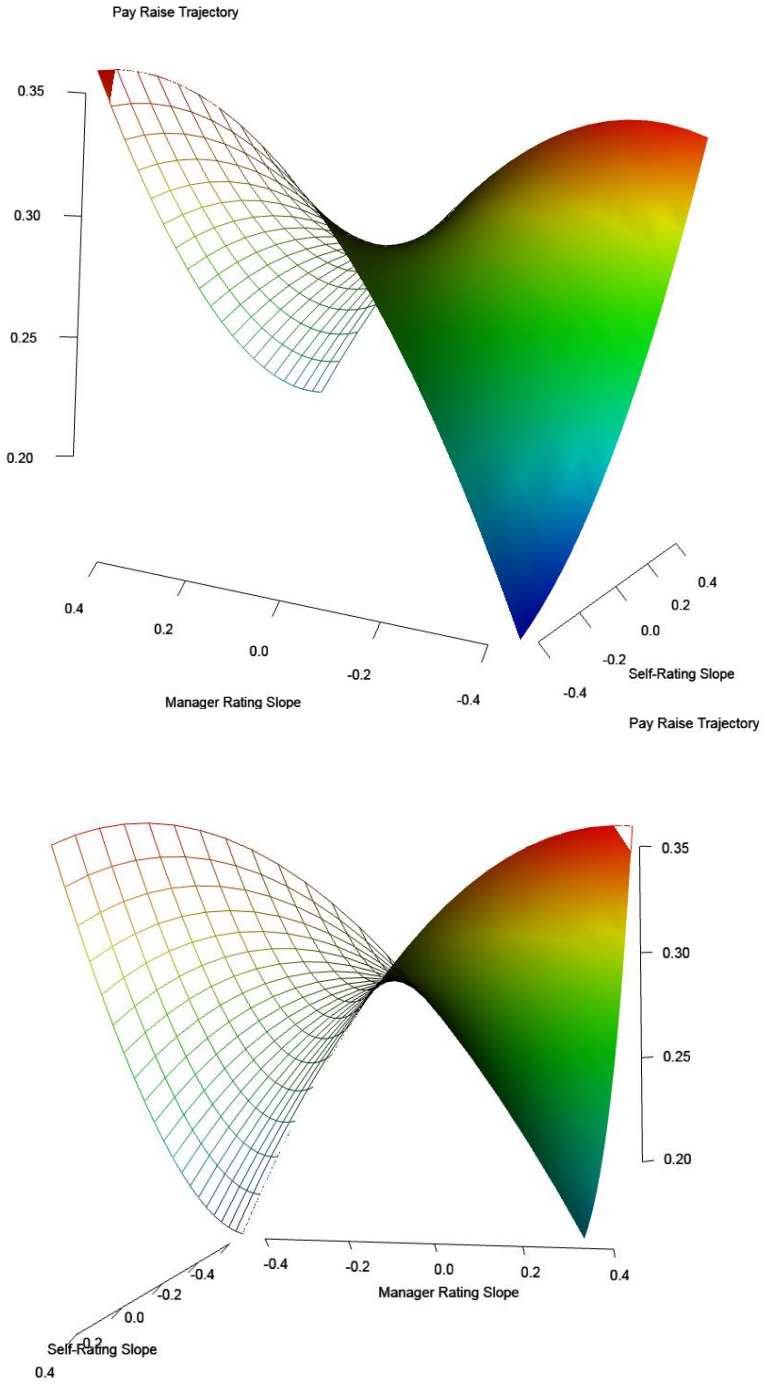

Figure 12: Response surface for pay raise trajectories (in # raises per year) predicted by rating trajectories.

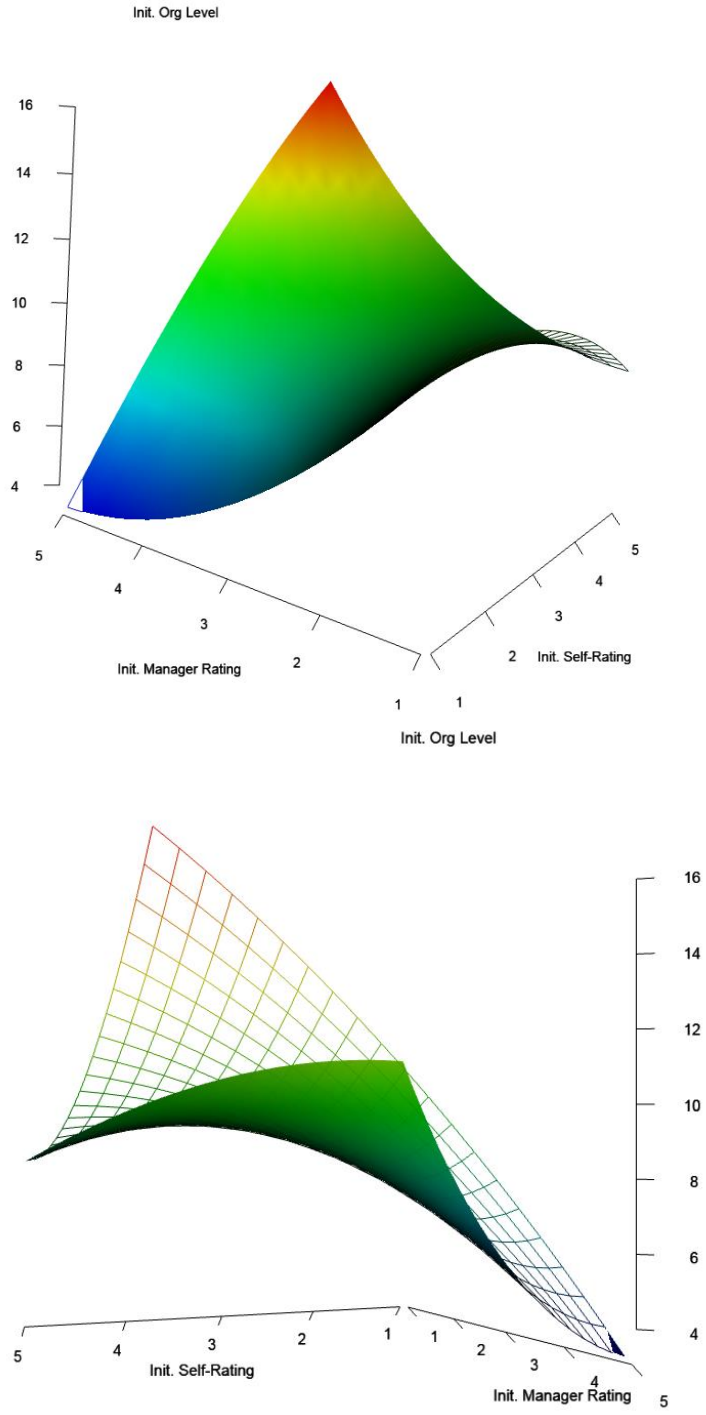

Figure 13: Response surface for initial organizational level predicted by initial performance ratings.

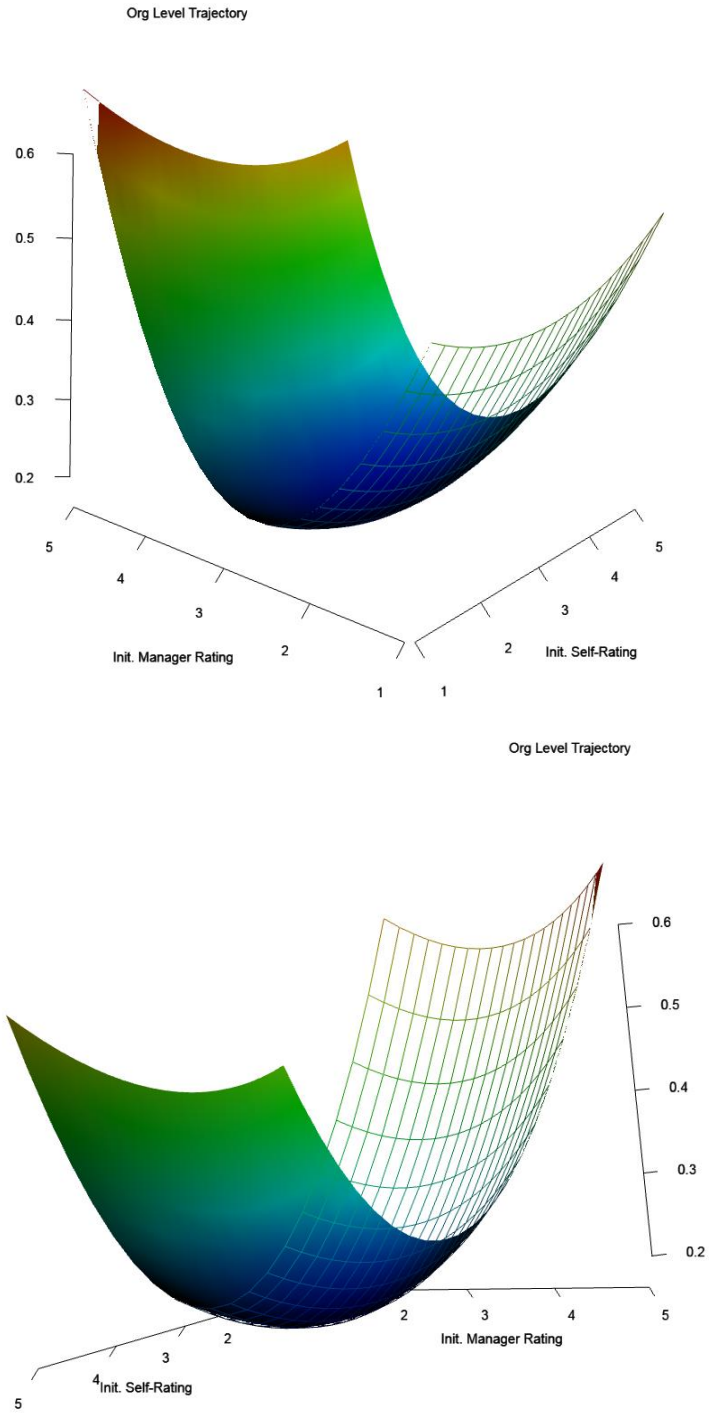

Figure 14: Response surface for organizational level trajectory predicted by initial performance ratings.

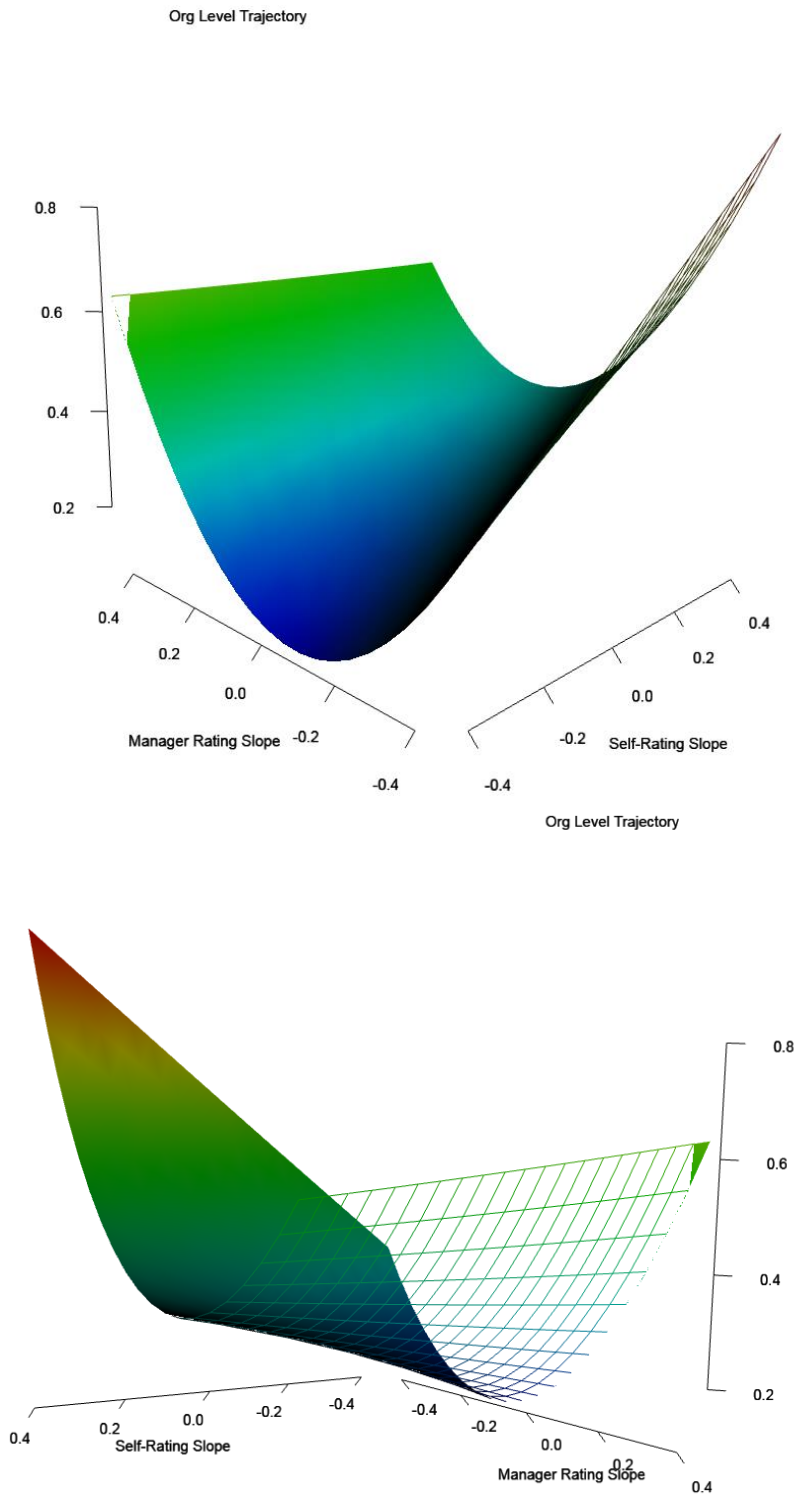

Figure 15: Response surface for organizational level trajectory predicted by performance ratings slopes.

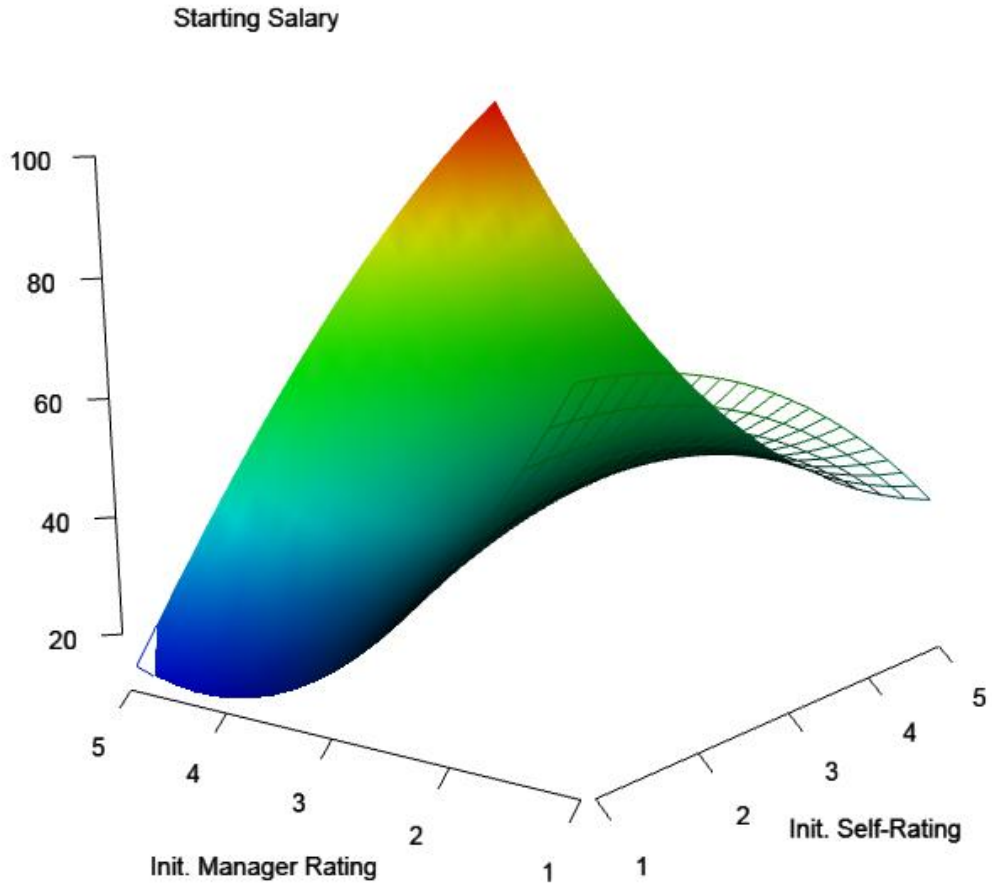

Figure 16: Response surface for initial salary (in $1k) predicted by initial performance ratings with control variables.

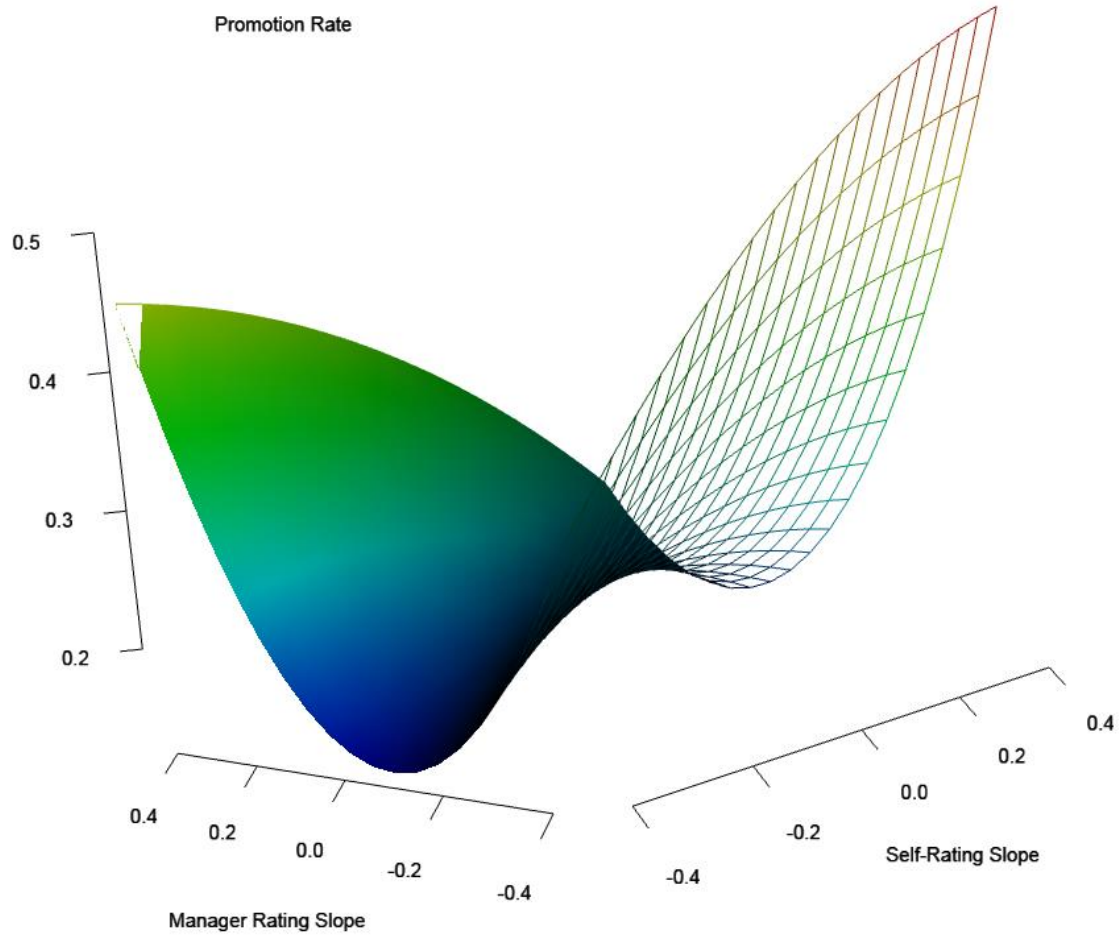

Figure 17: Response surface for promotion trajectories (expected # promotions per year) predicted by rating trajectories with control variables.

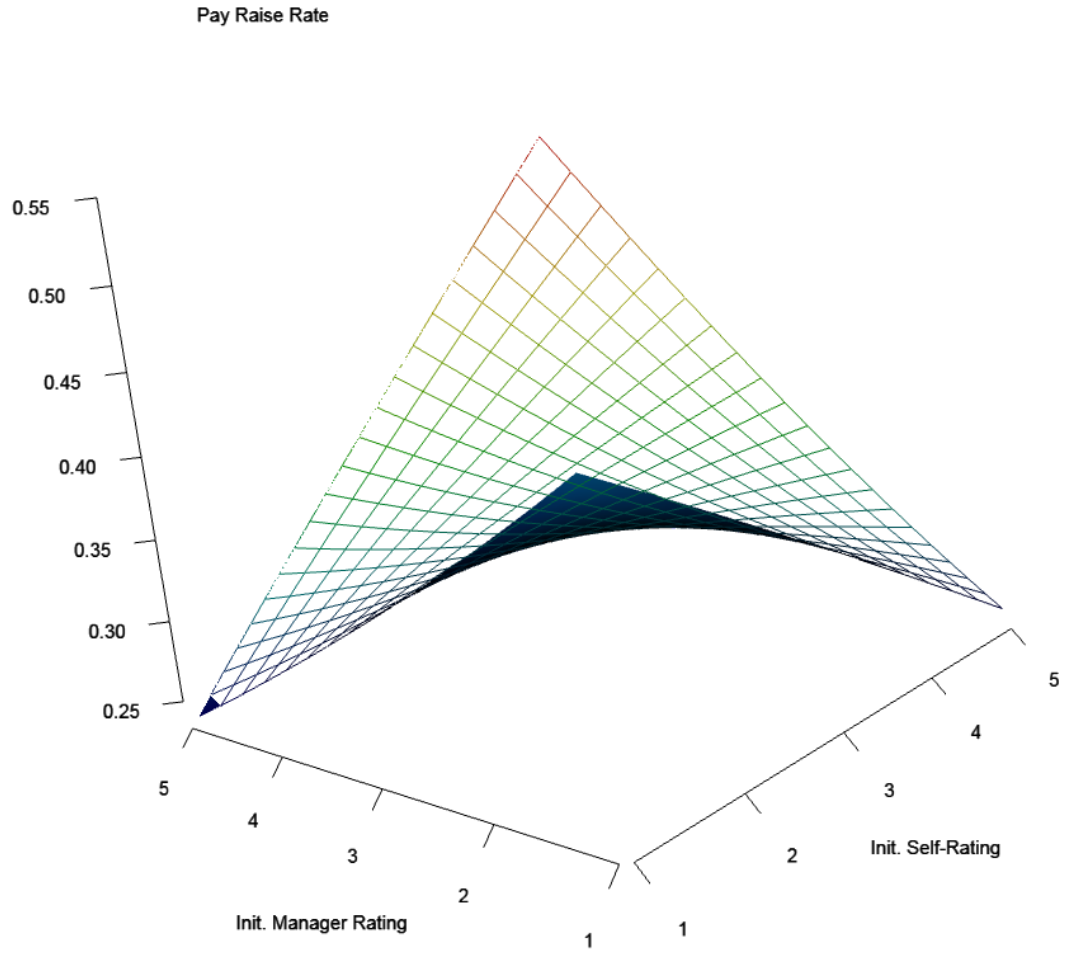

Figure 18: Response surface for pay raise trajectories (in # raises per year) predicted by initial performance ratings with control variables.

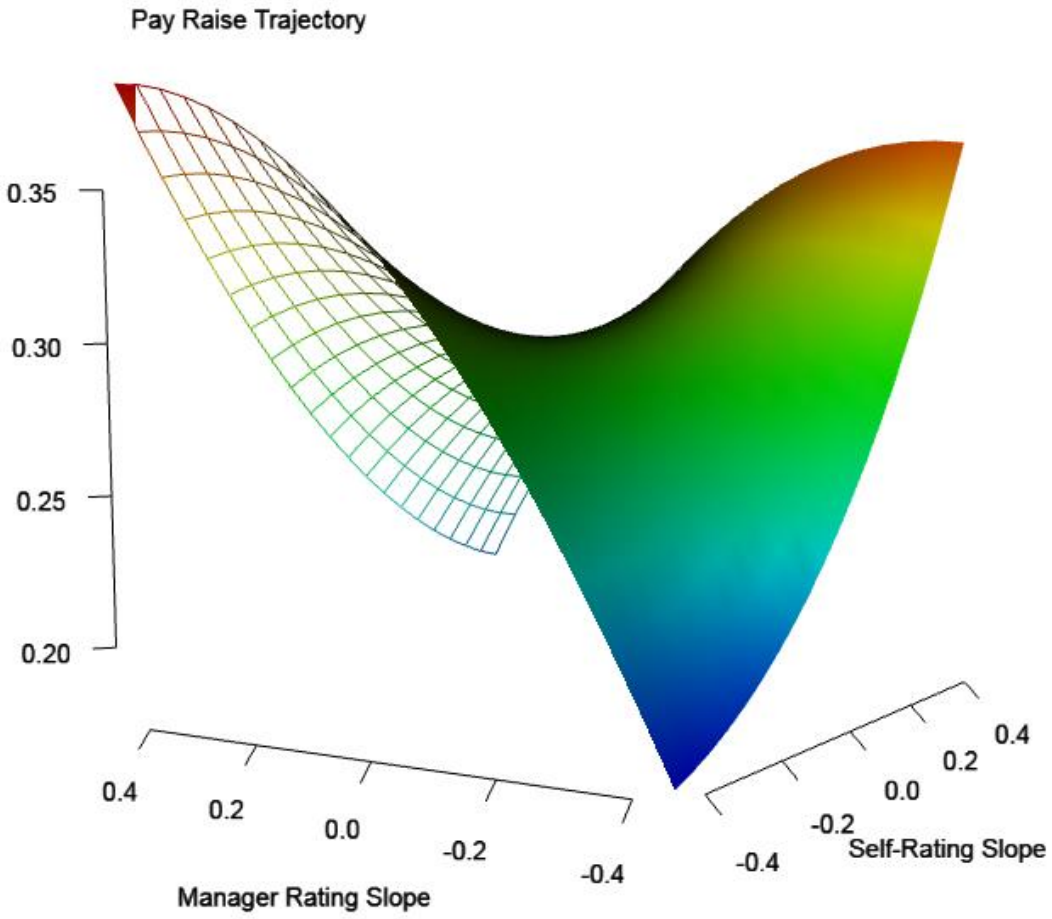

Figure 19: Response surface for pay raise trajectories (in # raises per year) predicted by rating trajectories with control variables.

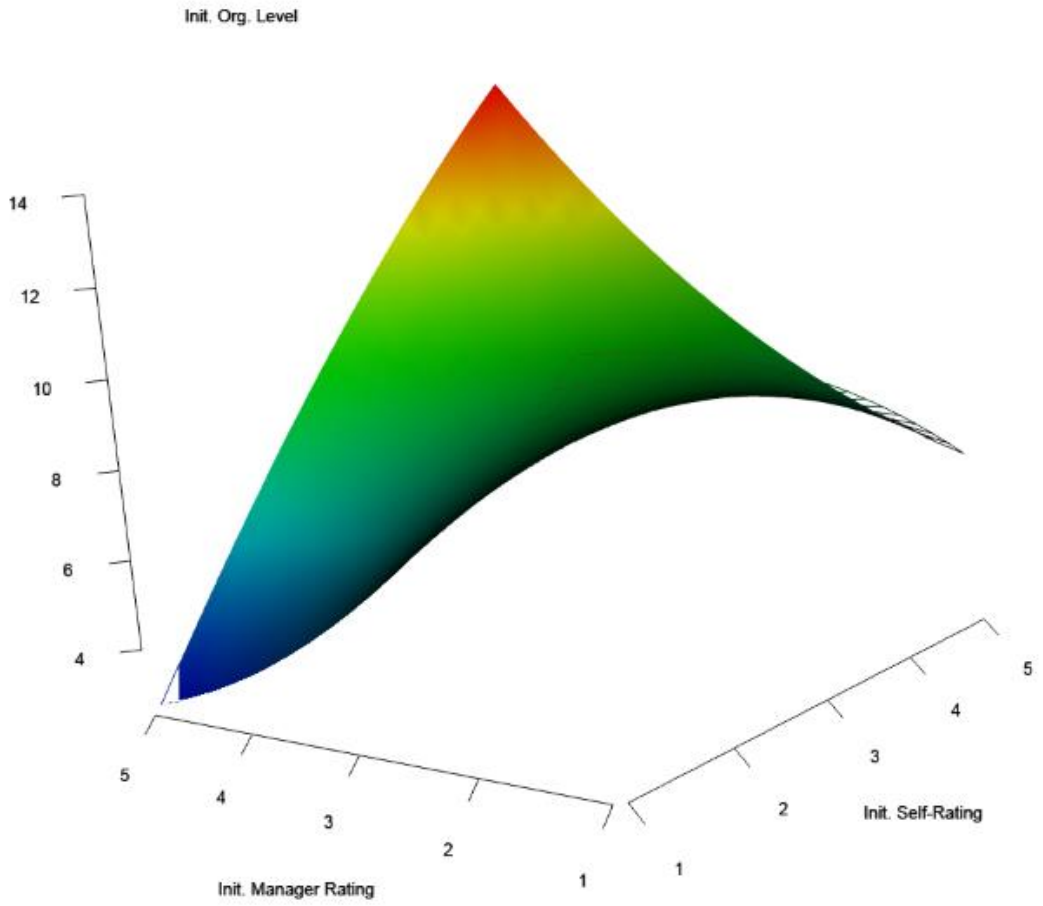

Figure 20: Response surface for initial organizational level predicted by initial performance ratings with control variables.

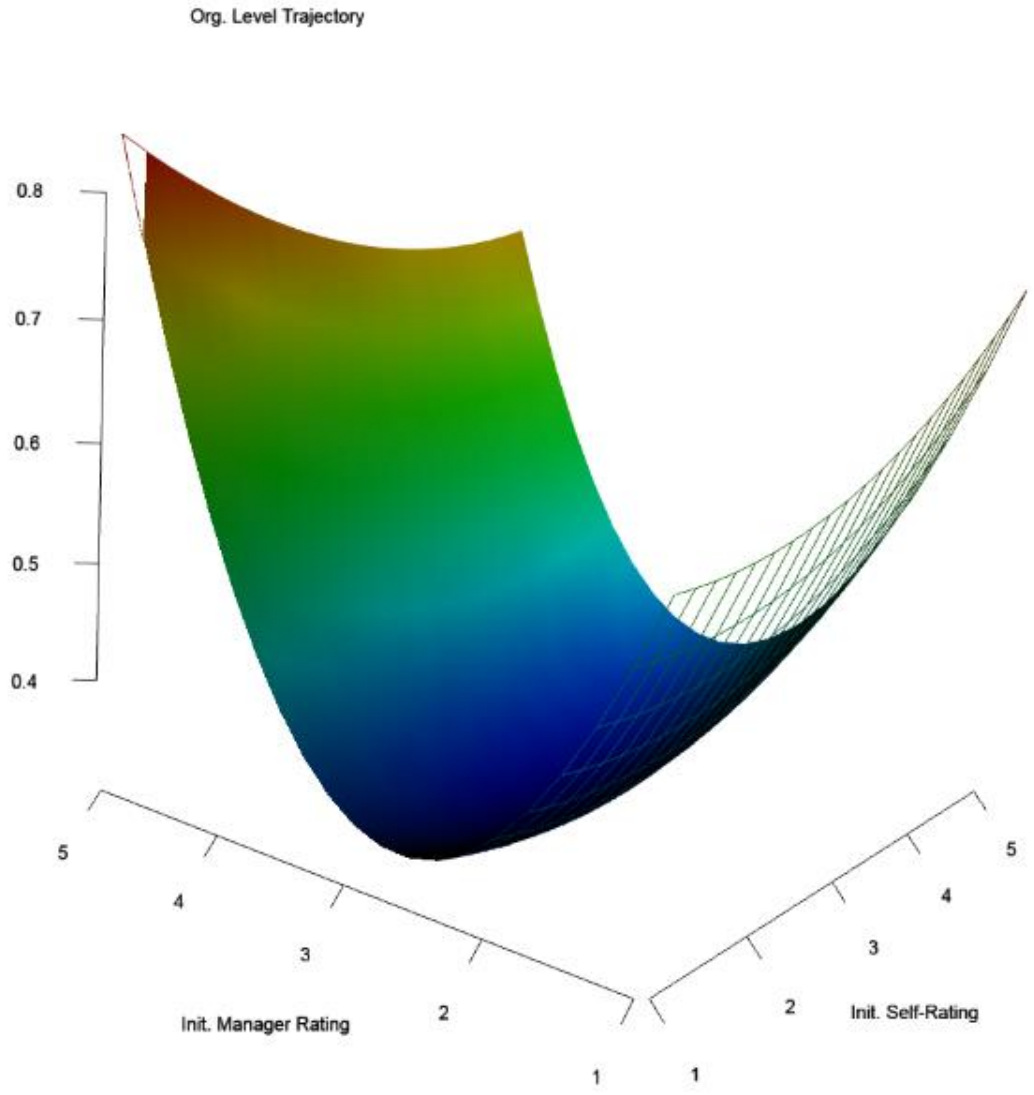

Figure 21: Response surface for organizational level trajectory predicted by initial performance ratings with control variables.

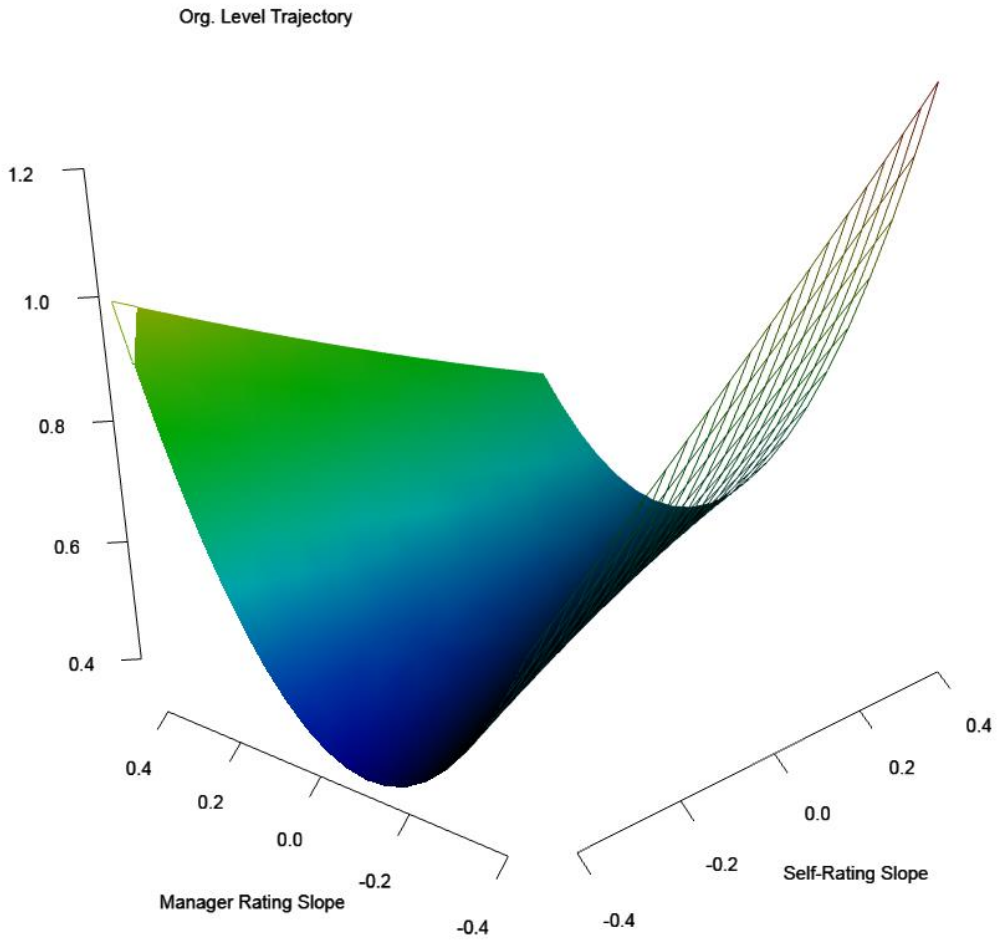

Figure 22: Response surface for organizational level trajectory predicted by performance ratings slopes with control variables.

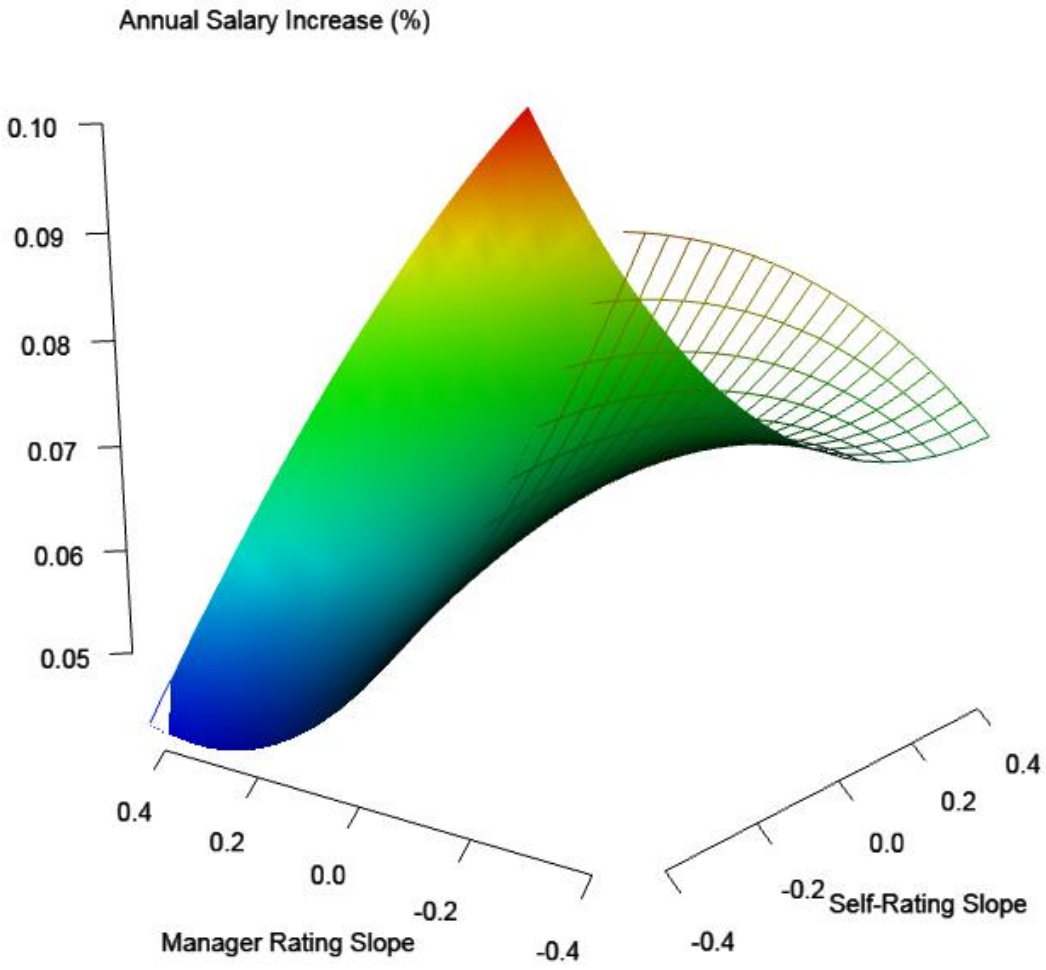

Figure 23: Response surface for salary trajectories predicted by longitudinal agreement with control variables.

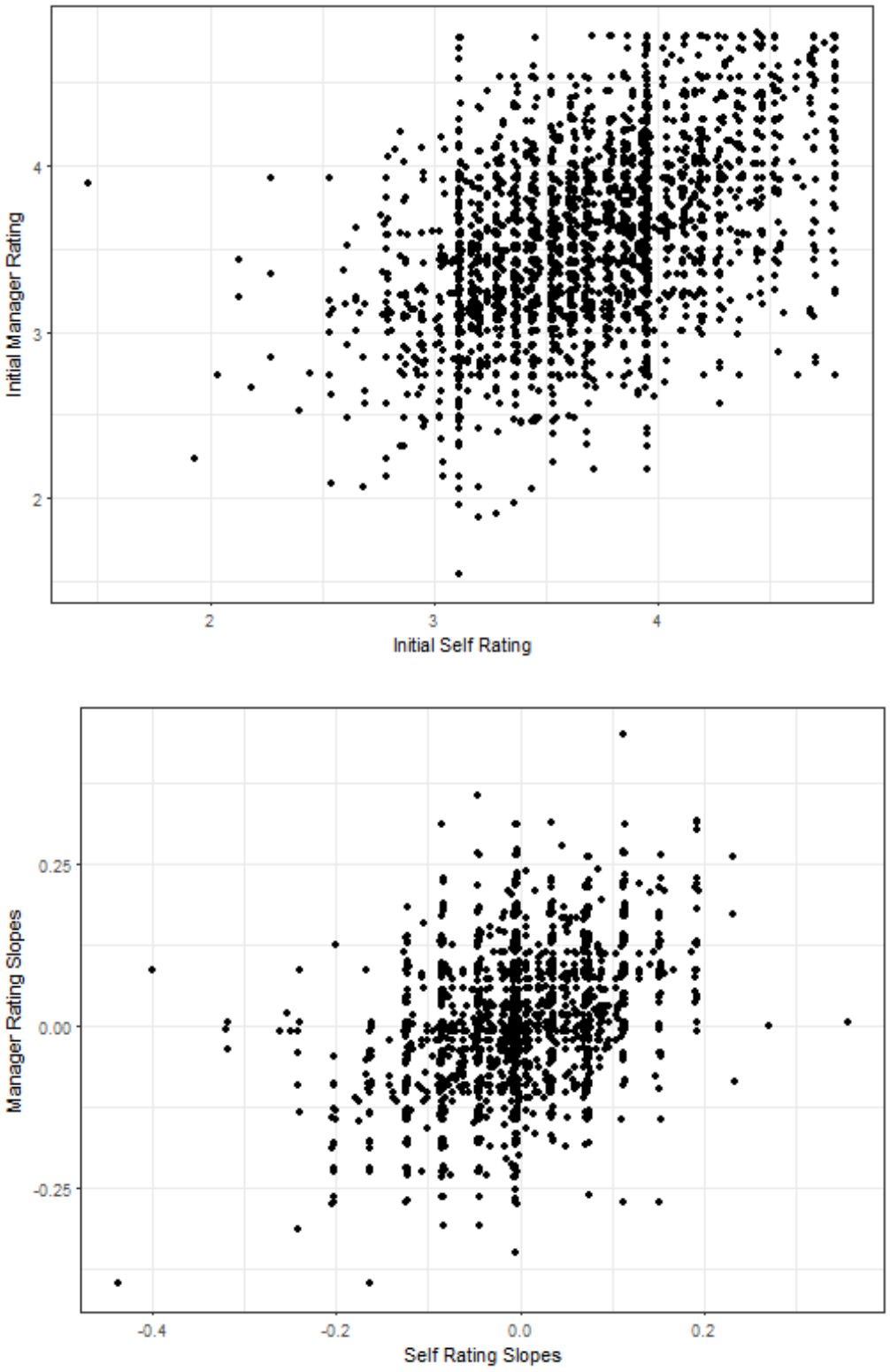

Figure 24: Scatterplots for polynomial regression data.

Table 1

Annual performance rating dimensions of the organization

Rating Level	*Criteria*
5 - Outstanding	• Performance consistently exceeds expectations even under challenging circumstances; sets and accomplishes stretch goals • Possesses, demonstrates, and leverages higher level capabilities in addition to competencies, skillsets, or knowledge critical for effective job performance. • Performance and capabilities are significantly higher than most peers in similar jobs/levels
4 – More than fully contributing	• Performance consistently meets and often exceeds expectations; sets and accomplishes most stretch goals • Possesses and demonstrates the competencies, skillsets, or knowledge critical for effective job performance and is actively developing higher capabilities. • Performance and capabilities are higher than most peers in similar jobs/levels
3 – Fully contributing	• Performance consistently meets expectations and achieves agreed upon goals/objectives including some stretch goals. • Possesses and demonstrates the competencies, skillsets, or knowledge critical for effective job performance and is continuously upgrading personal capability. • Performance and capabilities match most peers in similar jobs/levels
2 – Needs some improvement	• Performance meets some, but not all, expectations and achieves some, but not all, goals/objectives; making some progress at improving performance but still requires frequent guidance. • Lacks some of the competencies, skillsets, or knowledge critical for effective job performance. • Performance and capabilities are lower than most peers in similar jobs/levels
1 – Needs significant improvement	• Performance is below expectations and usually does not achieve agreed upon goals/objectives; not making required progress at improving performance despite repeated feedback and guidance. • Lacks many of the competencies, skillsets, or knowledge critical for effective job performance. • Performance and capabilities are significantly lower than most peers in similar jobs/levels

Table 2

Competency model of the organization—names and descriptions

Competency	*Description*
Demonstrate Agility	Responds resourcefully, flexibly, and positively when faced with new challenges and demands, moving forward productively under conditions of change or uncertainty. Learns from setbacks or mistakes and quickly bounces back. Deals effectively with ambiguity.
Make Insightful Decisions	Analyzes both problems and opportunities and their impact on the business, customers and other key stakeholders. Integrates information, data, guidelines, and requirements from different sources to evaluate alternatives and make effective, timely and well-reasoned decisions.
Act Boldly	Challenges the status quo. Tackles tough assignments, proactively holds courageous conversations, steps forward to address difficult issues with transparency, and supports others who do so. Capably balances risk with reward. Doesn't hold back on anything that needs to be said. Maintains relationships while acting boldly.
Lead & Embrace Change	Initiates and leads change to ensure continuous improvement and make the organization successful. Responds resourcefully and constructively to new opportunities to learn and grow and new ways of getting work done.
Act Strategically	Aligns the capabilities and strategies of the organization to capture emerging trends, address competitive threats, meet market needs, and provide value to customers. Works to enhance organizational value and create competitive advantage.
Innovate to Grow	Generates and champions new ideas, approaches and initiatives, and creates an environment that nurtures innovation. Supports business transformation by encouraging new ways of looking at problems, processes or solutions.
Engage & Include	Builds relationships inside and outside of the organization that enhance the levels of cooperation, collaboration and trust. Fosters a culture that makes people feel valued and respected, appreciates diverse opinions, and even in difficult or tense circumstances builds trust and enhances relationships. Promotes a free and timely flow of high quality information and ideas across the organization.
Influence & Inspire	Articulates a compelling rationale that inspires commitment to a point of view or plan of action. Instills and sustains energy and optimism, helping others to envision a greater sense of what is possible. Models personal commitment.

Build Talent	Attracts, develops, manages, and retains critical and diverse talent. Provides coaching, feedback, and support, and shares best practices, effectively and constructively enabling individuals to achieve high performance. Differentiates and recognizes outstanding performance and deals constructively with underperformance in a timely manner.
Demonstrate Integrity in Products, Processes & Relationships	Demonstrates principled leadership and sound business ethics; shows consistency among principles, values, and behavior; builds trust with others through own authenticity.
Execute with Focus & Accountability	Ensures work performance and accountability, demonstrating and fostering a sense of urgency and focus. Has a "can-do" spirit, a sense of ownership, and a strong commitment to achieving goals, meeting customer requirements and organizational success.

Table 3

Illustration of the possible longitudinal agreement categories

		Intercept Difference (Manger – Self)		
		Negative	None	Positive
Slope Difference (Manager – Self)	Positive	Reformed Over-rater	Deflated Agreer	Serial Under-rater
	None	Consistent Over-rater	Consistent Agreer	Consistent Under-rater
	Negative	Serial Over-rater	Inflated Agreer	Reformed Under-rater

Table 4

Descriptive statistics for appraisal ratings by year

Year	Self-Ratings		Manager Ratings		Manager – Self	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
2011	3.635	0.686	3.534	0.807	-0.106	0.813
2012	3.671	0.706	3.575	0.774	-0.100	0.771
2013	3.660	0.703	3.548	0.759	-0.114	0.792
2014	3.615	0.676	3.505	0.744	-0.110	0.759
2015	3.540	0.770	3.494	0.736	-0.074	0.755
0	3.641	0.676	3.536	0.806	-0.105	0.812
1	3.675	0.696	3.577	0.772	-0.098	0.769
2	3.666	0.691	3.548	0.758	-0.118	0.780
3	3.641	0.664	3.535	0.756	-0.106	0.759
4	3.599	0.698	3.540	0.742	-0.059	0.742

Table 5

Agreement statistics by performance year

Year	ρ	κ (unweighted)	κ (linear weights)	κ (squared weights)
2011	0.421	0.284	0.336	0.400
2012	0.454	0.273	0.349	0.448
2013	0.420	0.283	0.337	0.403
2014	0.432	0.299	0.353	0.425
2015	0.469	0.331	0.384	0.449
0	0.421	0.284	0.336	0.401
1	0.454	0.273	0.349	0.448
2	0.427	0.284	0.341	0.416
3	0.435	0.307	0.359	0.426
4	0.489	0.347	0.400	0.478

ρ Spearman's rho for rank-order agreement
k (unweighted) Cohen's kappa where all levels of disagreement are treated equally
k (linear weights) Cohen's kappa where increased disagreement is penalized but at a constant level
k (squared weights) Cohen's kappa where increased disagreement is more severely penalized

Table 6

Summaries of models fit to salary data (N = 3,403)

Criterion:	Salary Intercepts[a]		Salary Trajectories[b]		Salary Trajectories	
Predictor:	Ratings Intercepts		Ratings Intercepts		Ratings Slopes	
Intercept	-$16,132	$149,965	0.006	-0.069	0.051	0.051
Self	$10,814	$2,599	-0.003	0.004	0.029	0.029
Manager	$13,294	-$71,228	0.015	0.051	-0.008	-0.007
Self^2		-$2,737		0.001		-0.015
Manager^2		$7,755		-0.003		0.045
Self $\times$ Mgr		$7,753		-0.004		0.090
R^2(adj)	0.067	0.075*	0.050	0.051*	0.003	0.005

[a] model coefficients predicting salary intercepts are interpreted as dollars
[b] model coefficients predicting salary slopes are interpreted as % change in salary per year
* difference in model fit $p < 0.05$

Table 7

Summaries of models fit to promotions data (N = 3,403)

Criterion:	Promotion Trajectories[a]		Promotions Trajectories	
Predictor:	Ratings Intercepts		Ratings Slopes	
Intercept	-0.050	-0.154	0.170	0.164
Self	0.003	0.031	0.120	0.119
Manager	0.034	0.064	-0.070	-0.078
$Self^2$		-0.006		-0.118
$Manager^2$		-0.006		0.944
Self × Mgr		0.004		-0.729
R^2(adj)	0.030	0.030	0.001	0.003*

[a] model coefficients predicting promotion slopes are interpreted as ½ expected # promotions per year
* difference in model fit $p < 0.05$

Table 8

Summaries of models fit to pay raise data (N = 3,403)

Criterion:	Pay Raise Trajectories[a]		Pay Raise Trajectories	
Predictor:	Ratings Intercepts		Ratings Slopes	
Intercept	0.422	0.769	0.289	0.291
Self	-0.017	-0.129	0.016	0.018
Manager	-0.020	-0.099	0.033	0.030
Self^2		0.003		0.161
Manager^2		-0.002		-0.276
Self $\times$ Mgr		0.025		-0.425
R^2(adj)	0.024	0.027*	0.001	0.004*

[a] model coefficients predicting pay raise slopes are interpreted as ½ expected # raises per year
* difference in model fit $p < 0.05$

Table 9

Summaries of models fit to organizational level data (N = 3,401)

Criterion:	Org. Level Intercepts		Org. Level Trajectories		Org Level Trajectories	
Predictor:	Ratings Intercepts		Ratings Intercepts		Ratings Slopes	
Intercept	-0.041	16.779	-0.266	0.906	0.241	0.226
Self	1.546	-0.371	0.018	-0.060	0.277	0.282
Manager	1.314	-6.116	0.123	-0.456	-0.071	-0.082
$Self^2$		-0.248		0.015		0.013
$Manager^2$		0.500		0.086		1.935
Self × Mgr		1.032		-0.010		-0.858
R^2(adj)	0.073	0.078*	0.017	0.020*	0.001	0.002*

* difference in model fit $p < 0.05$

Table 10

Summaries of models fit to salary data with control variables (N = 3,403)

Criterion:	Salary Intercepts[a]		Salary Trajectories[b]		Salary Trajectories	
Predictor:	Ratings Intercepts		Ratings Intercepts		Ratings Slopes	
Intercept	-$28,898	$129,355	0.013	-0.054	0.060	0.059
Self	$10,039	-$2,856	-0.002	0.012	0.019	0.019
Manager	$14,097	-$51,551	0.015	0.039	-0.010	-0.009
Self^2		-$2,137		< 0.001		-0.024
Manager^2		$5,062		< 0.001		0.079
Self $\times$ Mgr		$7,873		-0.005		0.101
Ethnicity[c]						
Native	$4,842	$3,617	0.003	0.004	0.007	0.007
Asian	$8,009	$8,595	-0.001	-0.001	-0.001	-0.001
Black	$6,909	$7,011	-0.001	-0.001	-0.003	-0.002
Hispanic	-$383	-$11	0.001	0.001	< 0.001	0.001
Pacific Is.	-$5,303	$5,599	0.022	0.022	0.024	0.020
$\geq$ 2 races	$3,466	$3,404	-0.002	-0.003	-0.004	-0.003
Gender[d]	-$22,025	-$21,756	-0.001	-0.002	-0.001	-0.001
Tenure	$689	$682	-0.001	-0.001	-0.001	-0.001
R^2(adj)	0.165	0.170*	0.133	0.134	0.082	0.084*

[a] model coefficients predicting salary intercepts are interpreted as dollars
[b] model coefficients predicting salary slopes are interpreted as % change in salary per year
[c] reference group is employee ethnicity = white
[d] reference group is employee gender = male
* difference in model fit $p < 0.05$

Table 11

Summaries of models fit to promotions data with control variables (N = 3,403)

Criterion:	Promotion Trajectories[a]		Promotions Trajectories	
Predictor:	Ratings Intercepts		Ratings Slopes	
Intercept	-0.069	-0.212	0.214	0.203
Self	0.009	0.098	0.067	0.064
Manager	0.070	0.059	-0.082	-0.089
Self^2		-0.016		-0.200
Manager^2		-0.003		1.096
Self $\times$ Mgr		0.008		-0.615
Ethnicity[b]				
Native	-0.015	-0.014	0.006	0.010
Asian	-0.005	-0.005	-0.002	-0.001
Black	0.038	0.037	0.030	0.032
Hispanic	0.038	0.038	0.034	0.034
Pacific Is.	-0.001	-0.003	0.002	-0.027
≥ 2	-0.032	-0.031	-0.035	-0.033
Gender[c]	-0.004	-0.004	-0.002	-0.003
Tenure	-0.006	-0.006	-0.006	-0.006
R^2(adj)	0.082	0.082	0.052	0.056*

[a] model coefficients predicting promotion slopes are interpreted as ½ expected # promotions per year
[b] reference group is employee ethnicity = white
[c] reference group is employee gender = male
* difference in model fit $p < 0.05$

Table 12

Summaries of models fit to pay raise data with control variables (N = 3,403)

Criterion:	Pay Raise Trajectories[a]		Pay Raise Trajectories	
Predictor:	Ratings Intercepts		Ratings Slopes	
Intercept	0.438	0.777	0.301	0.304
Self	-0.017	-0.121	0.002	0.005
Manager	-0.019	-0.105	0.031	0.027
Self^2		0.002		0.164
Manager^2		-0.001		-0.238
Self $\times$ Mgr		0.024		-0.451
Ethnicity[b]				
Native	0.019	0.015	0.008	0.009
Asian	0.003	0.005	0.001	0.001
Black	-0.005	-0.004	-0.001	-0.002
Hispanic	0.007	0.007	0.008	0.007
Pacific Is.	-0.006	-0.006	-0.004	0.010
≥ 2	0.017	0.017	0.017	0.015
Gender[c]	-0.013	-0.013	-0.014	-0.014
Tenure	-0.001	-0.001	-0.001	-0.001
R^2(adj)	0.034	0.037*	0.011	0.014*

[a] model coefficients predicting pay raise slopes are interpreted as expected # raises per year
[b] reference group is employee ethnicity = white
[c] reference group is employee gender = male
* difference in model fit $p < 0.05$

Table 13

Summaries of models fit to organizational level data with control variables (N = 3,401)

Criterion:	Org. Level Intercepts		Org. Level Trajectories		Org Level Trajectories	
Predictor:	Ratings Intercepts		Ratings Intercepts		Ratings Slopes	
Intercept	-0.319	15.43	-0.189	0.941	0.329	0.314
Self	1.495	-0.990	0.019	-0.014	0.185	0.190
Manager	1.339	-4.907	0.127	-0.476	-0.081	-0.096
Self^2		-0.190		0.012		0.054
Manager^2		0.313		0.092		2.190
Self $\times$ Mgr		1.073		-0.017		-1.151
Ethnicity[a]						
Native	0.503	0.333	0.162	0.158	0.210	0.218
Asian	0.744	0.806	-0.064	-0.061	-0.058	-0.055
Black	1.342	1.365	0.019	0.019	0.005	0.009
Hispanic	-0.244	-0.202	0.031	0.032	0.023	0.023
Pacific Is.	-0.268	0.302	0.159	0.162	0.174	0.115
≥ 2	1.326	1.322	-0.014	-0.016	-0.021	-0.017
Gender[b]	-0.541	-0.516	-0.102	-0.099	-0.098	-0.099
Tenure	0.088	0.087	-0.008	-0.008	-0.007	-0.008
R^2(adj)	0.100	0.104*	0.039	0.042*	0.021	0.023*

[b] reference group is employee ethnicity = white
[c] reference group is employee gender = male
* difference in model fit $p < 0.05$

Table 14

Summary of conclusions supported by all response surface analyses

Outcome	Initial Agreement	Longitudinal Agreement	Initial Agreement (with controls)	Longitudinal Agreement (with controls)
Salary				
Initial	B	-	B	-
Trajectory	B	-	-	C
Promotion Rate	-	D	-	D
Pay Raise Rate	NA	D	NA	D
Org. Level				
Initial	B	-	B	-
Trajectory	B	C, D	B	C,D

A: agreement is unilaterally better
B: the effect of agreement is a function of performance level
C: agreement about growth is better than agreement about decline
D: convergence in ratings is better

Table 15

Survival analysis results from managerial change data

Agreement Category	*N*	*Observed*	*Expected*	$\frac{(O-E)^2}{E}$ [a]	$\frac{(O-E)^2}{V}$ [b]
Consistent Agreer	41	18	20.4	0.286	0.29
Consistent Over-rater	181	88	87.7	> 0.001	> 0.01
Consistent Under-rater	94	33	55.4	9.084	9.54
Deflated Agreer	39	17	11.0	3.265	3.33
Inflated Agreer	90	44	47.2	0.218	0.23
Reformed Over-rater	1368	667	642.4	0.944	1.60
Reformed Under-rater	705	310	343.0	3.178	4.10
Serial Over-rater	607	297	248.6	9.428	11.30
Serial Under-rater	370	133	151.2	2.194	2.45

[a] analogous to a variance; the magnitude of the deviance between observed and expected
[b] analogous to a χ^2; tests against the null hypothesis that expected is equal to observed; any value > 3.84 is significant

Table 16

Models applied to appraisal data to test for feedback effects

	Baseline	*Δ Intercepts*	*Δ Slope*	*Δ Int. & Slope*
Self Ratings				
Fixed Effects				
Intercept	3.797	3.794	3.797	3.797
Time	0.004	0.010	0.003	0.007
Participation		-0.026		-0.037
Time(since 360)			0.002	0.011
Variance components				
Within person	0.229	0.229	0.229	0.229
In initial status	0.285	0.285	0.230	0.286
In rate of change	0.017	0.017	0.017	0.017
Goodness-of-fit				
Deviance	1500.0	1499.8	1500.0	1499.7
AIC	1512.0	1513.8	1514.0	1515.7
BIC	1540.3	1546.8	1547.0	1553.4
Manager Ratings				
Fixed Effects				
Intercept	3.786	3.763	3.757	3.753
Time	-0.002	0.064	0.051	0.077
Participation		-0.253		-0.208
Time(since 360)			-0.096	-0.044
Variance components				
Within person	0.364	0.358	0.359	0.357
In initial status	0.306	0.294	0.299	0.294
In rate of change	0.014	0.015	0.017	0.016
Goodness-of-fit				
Deviance	1787.4	1775.3	1780.3	1774.1
AIC	1799.4	1789.3	1794.3	1790.1
BIC	1827.7	1822.3	1827.3	1827.8

Table 17

Intra-class correlations for models fit to longitudinal data

Model	*ICC*
Self rating ~ Year	0.561
Manager rating ~ Year	0.549
Salary ~ Year	0.986
Promotion rate ~ Year	0.294
Organization level ~ Year	0.917
Pay raise rate ~ Year	0.461

Appendix A

Items on the multisource feedback instrument:

1. Knows who to involve in decision making.
2. Follows through on commitments.
3. Communicates progress, problems, and information to key stakeholders in a timely manner.
4. Develops new and innovative approaches.
5. Anticipates and deals effectively with resistance to change.
6. Seeks performance feedback and acts on it to improve self.
7. Develops and implements plans to effectively manage change while simultaneously maintaining operating effectiveness.
8. Works with employees to develop their skills; provides coaching and feedback.
9. Builds positive and effective working relationships both within the organization and with external partners and customers.
10. Fosters an environment that promotes continuous improvement and desired outcomes of change.
11. Approaches problems and opportunities from a broad company perspective.
12. Willingly takes on challenging assignments.
13. Treats everyone with dignity and respect, regardless of their background, position, or situation.
14. Speaks up and is willing to take an unpopular stand.
15. Understands the financial implications of decisions and actions.
16. Can think on his/her feet.
17. Conducts business with honesty and integrity.
18. Addresses performance problems in a timely manner.
19. Makes good decisions in a timely manner.
20. Evaluates and pursues initiatives, investments, and opportunities based on their fit with broader company strategies.
21. Ensures others understand the way their work supports broader organizational strategies.
22. Takes action now to position the organization for future success.
23. Is good at generating long-term and strategic ideas to drive future business.
24. Is resilient, especially in new or challenging situations.
25. Gives honest answers to questions and challenges.
26. Keeps others informed.
27. Willing to challenge upwardly (with his/her manager, other leaders).
28. Honors agreements made with others.
29. Is open in communications with others, even when it's easier not to be.
30. Creates or sponsors systems and processes that encourage creativity and innovation.

31. Is open to others' ideas and recommendations.
32. Learns from mistakes.
33. Takes action to build a diverse and respectful team.
34. Remains calm and controlled in difficult situations.
35. Collaborates effectively across all groups.
36. Motivates and inspires the team to work toward the organization's mission and values.
37. Is resourceful and figures out how to get things done in spite of obstacles.
38. Thinks "outside the box" and comes up with new ways of doing things.
39. Communicates clearly and concisely.
40. Takes a big picture perspective.
41. Rewards and recognizes outstanding behavior.
42. Willing to take calculated risks.
43. Holds people accountable.
44. Leverages leading-edge technologies, processes, tools, and practices to contribute to the organization's future.
45. Makes decisions that are the best for the overall company.
46. Takes responsibility for actions, including decisions made and mistakes.
47. Manages his/her time effectively.
48. Rewards and recognizes creativity and innovation.
49. Helps others to seek feedback from and actively manage stakeholders.
50. Ensures that products, processes, and relationships are sound and ethical.
51. Effectively balances honesty with compassion.
52. Demonstrates consistency among principles, values, and behavior.